Title page

A systematic review of measurement properties of patient reported outcome measures in psoriatic arthritis: a GRAPPA-OMERACT initiative

Pil Højgaard[a,b], Louise Klokker[a], Ana-Maria Orbai[c], Kim Holmsted[a], Else M. Bartels[a], Ying Ying Leung[d], Niti Goel[e], Maarten de Wit[f], Dafna D. Gladman[g], Philip Mease[h], Lene Dreyer[a,b], Lars E. Kristensen[a], Oliver FitzGerald[i], William Tillett[j], Laure Gossec[k], Philip Helliwell[l], Vibeke Strand[m], Alexis Ogdie[n], Caroline Terwee[o], Robin Christensen[a].

a The Parker Institute, Copenhagen University Hospital, Bispebjerg and Frederiksberg , Copenhagen, Denmark pil.hoejgaard.01@regionh.dk, Louise.Klokker.Madsen@regionh.dk, holmsted73@yahoo.com, else.marie.bartels@regionh.dk, lars.erik.kristensen@regionh.dk, Robin.christensen@regionh.dk

b Department of Rheumatology, Rigshospitalet, Gentofte Hospital, Kildegaardsvej 28, 2900 Hellerup, Denmark Lene.Dreyer@regionh.dk

c Johns Hopkins University School of Medicine, Division of Rheumatology, Baltimore Maryland, USA aorbai1@jhmi.edu

d Department of Rheumatology and Immunology, Singapore General Hospital, Singapore (20 College Road, the Academia, S169856, Singapore) katyccc@hotmail.com

e Division of Rheumatology, Duke University School of Medicine; Advisory Services, QuintilesIMS; Patient Research Partner; Durham, NC, niti.goel@quintilesims.com

f VU University Amsterdam, Department of Medical Humanities, Amsterdam, Netherlands martinusdewit@hotmail.com

g Devision of Rheumatology and Krembil Research Institute, University Health Network, Toronto Western Hospital, 399 Bathurst Street, 1E-410B, Toronto, Ontario, M5T 2S8 dafna.gladman@utoronto.ca

h Division of Rheumatology Clinical Research, Swedish Medical Center, Seattle, WA, USA, pmease@philipmease.com

i Department of Rheumatology, St Vincent's University Hospital and Conway Institute for Biomolecular Research, University College Dublin, IRELAND. oliver.fitzgerald@ucd.ie

j Department of Rheumatology, Royal National Hospital for Rheumatic Diseases, Bath, UK & Pharmacy and Pharmacology, University of Bath, Bath, UK. w.tillett@nhs.net

k Sorbonne Universités, UPMC Univ Paris 06, GRC-08, Institut Pierre Louis d'Epidémiologie et de Santé Publique, Paris, France; Pitie-Salpétrière Hospital, AP-HP, Rheumatology department, Paris, France postal adress 47-83 Bd de l'Hopital 75013 Paris France laure.gossec@aphp.fr

l Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, 2nd Floor, Chapel Allerton Hospital, Harehills Lane Leeds, LS7 4SA p.helliwell@leeds.ac.uk
m Division of Immunology/Rheumatology, Stanford University, Palo Alto, CA, USA. vstrand@stanford.edu
n Division of Rheumatology, Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. Alexis.Ogdie @uphs.upenn.edu
o VU University Medical Center, Department of Epidemiology and Biostatistics and the and Amsterdam Public Health research institute, Amsterdam, the Netherlands cb.terwee@vumc.nl

Correspondence to:

Robin Christensen, BSc, MSc, PhD (Biostatistician)

Professor of clinical epidemiology, adj.

Head of Musculoskeletal Statistics Unit,

The Parker Institute, Copenhagen University Hospital, Bispebjerg and Frederiksberg,

Nordre Fasanvej 57, DK-2000 Copenhagen F, Denmark

Phone: (+45) 3816 4165

Fax: (+45) 3816 4159

E-mail: Robin.Christensen@regionh.dk

Word count 4228 (including contribution and funding sections)

Content:

Supplementary files (AT THE END OF THE DOCUMENT)

ABSTRACT

Background: An updated psoriatic arthritis (PsA) core outcome set (COS) for randomized controlled trials (RCTs) was endorsed at the Outcome Measures in Rheumatology (OMERACT) meeting in 2016.

Objectives: Synthesize the evidence on measurement properties of patient reported outcome measures (PROMs) for PsA and thereby contribute to development of a PsA core outcome measurement set (COMS) as described by the OMERACT Filter 2.0.

Methods: A systematic literature search was performed in EMBASE, MEDLINE and PsycINFO on Jan 1st 2017 to identify full-text articles with an aim of assessing the measurement properties of PROMs in PsA. Two independent reviewers rated the quality of studies using the COnsensus based standards for the Selection of health Measurement INstruments (COSMIN) checklist, and performed a qualitative evidence synthesis.

Results: Fifty-five studies were included in the systematic review. Forty-four instruments and a total of 89 scales were analysed. PROMs measuring COS domains with at least fair quality evidence for good validity and reliability (and no evidence for poor properties) included the Stockerau Activity Score for PsA (German), Psoriasis Symptom Inventory, visual analogue scale for Patient Global, 36 Item Short Form Health Survey Physical Function subscale, Health Assessment Questionnaire Disability Index, Bath Ankylosing Spondylitis Functional Index, PsA Impact of Disease questionnaire, PsA Quality of Life questionnaire, VITACORA-19, Functional Assessment of Chronic Illness Therapy Fatigue scale and Social Role Participation Questionnaire.

Conclusions: At least one PROM with some evidence for aspects of validity and reliability was available for six of the eight mandatory domains of the PsA COS.

Keywords: psoriatic arthritis, OMERACT, COSMIN, patient reported outcome measures, measurement properties, systematic review

INTRODUCTION

Psoriatic arthritis (PsA) is a chronic inflammatory disease associated with a range of symptoms, co-morbidities and reduced health related quality of life.[1-3] Based on patients' and physicians' perspectives as well as recent research developments, the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA) together with the Outcome Measures in Rheumatology (OMERACT) international consensus effort developed an updated core outcome set (COS) for PsA[4], describing the outcomes (domains) that should be measured and reported in all randomized controlled trials. The updated PsA COS was endorsed in May 2016 by OMERACT and includes the following mandatory ('inner core') domains: Musculoskeletal (MSK) disease activity, Skin disease activity, Pain, Patient global, Physical function, Health Related Quality of Life (HRQoL), Fatigue and Systemic inflammation. Four other domains (Participation, Economic cost, Structural damage and Emotional well-being) were considered important but not mandatory (middle COS circle), and four domains (Sleep, Independence, Stiffness and Treatment burden) were placed in the "research agenda" (outer COS circle).[5]

The OMERACT Filter 2.0 provides guidelines for developing a core outcome measurement set (COMS) which comprises the appropriate instruments to assess each COS domain.[6] Great heterogeneity exists in instruments used for measuring the core domains of PsA, and several have been "borrowed" from other diseases without confirming their measurement properties in PsA.[7] Instruments should have evidence of validity, reliability and responsiveness as described in detail by the COnsensus based standards for the Selection of health Measurement INstruments organisation (COSMIN).[8] In addition, an instrument needs to be feasible and yield interpretable results.[9] These qualities are summarized by the original OMERACT Filter as 'Truth, Discrimination and Feasibility'.[10] As highlighted by the OMERACT Filter 2.0, the COS development was not influenced by considering *how* to measure the domains; neither the type of assessment nor the availability of specific instruments was taken into account. Development of the PsA COMS therefore implies that subsequently all available instruments per COS domain are identified,

evaluated and judged for overall applicability. To support this GRAPPA-OMERACT initiative, the objective of this systematic literature review was to synthesise the evidence for good measurement properties of patient reported outcomes measures (PROMS) in PsA and align instruments and COS domains.

METHODS

A protocol was uploaded to PROSPERO prior to initiation of the systematic review (PROSPERO: CRD42016032546). The review adheres to the COSMIN guidelines[11-13] and the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA-statement).[14]

Literature search

A research librarian (EMB) and the first-author (PH) performed a systematic search in MEDLINE via PubMed from 1966, EMBASE via OVID from 1974, and PsycINFO via OVID from 1806, all to 1 January 2017. The search was designed to identify all types of outcome measurement instruments in PsA. The search was limited to humans and consisted of two overall terms: *(1) Target population:* MeSH subheadings and free text words in title/abstract (ti/ab) were combined by the Boolean operator 'OR' to search for the target population (PsA) in the databases**;** *(2) Measurement properties:* Search filters have been developed to improve the search of studies on measurement properties in MEDLINE and EMBASE.[15] We used the highly sensitive filter validated for MEDLINE (sensitivity of 97.4%) and the filter for EMBASE optimized for this search. In PsycINFO only the target population was searched. The full search strategy is available in supplementary Table A.

Eligibility criteria

Per protocol, studies were considered eligible if published as full text articles in the English language with an aim of developing or assessing measurement properties of outcome measurements in PsA patients. However, for feasibility reasons and to ensure applicability of the COSMIN guidelines, it was subsequently decided to evaluate only patient reported instruments in this review, and allocate the assessment of the remaining instruments to parallel work streams. The stepwise eligibility and inclusion process is depicted in **Figure 1**. Studies evaluating instruments used solely for screening or diagnostic purposes were not

eligible. Only studies including ≥50% patients with PsA or reporting PsA subgroup results separately were included.

Selection of articles

PH eliminated duplicates and the remaining references were assessed for eligibility by two independent reviewers (PH, KH). Titles, abstracts and full-text articles (when appropriate) were reviewed and selection was performed by consensus with involvement of co-authors (RC, LK, EMB, A-MO) if needed. Additional studies identified by co-authors or reviews were considered for inclusion. Search results were handled by Reference Manager 12 (Thomson Reuters, USA).

Extraction of study characteristics and description of PROM characteristics

PH and KH independently extracted data on the characteristics of the studies (number, age and gender of participants, study setting and language). Characteristics of the PROMs (e.g., items, scoring, feasibility and availability) were obtained by PH from the questionnaires, background literature, user manuals or European League Against Rheumatism (EULAR) Outcome Measures Library[16] or by contacting authors/copyright holders.

Mapping the PROMs to corresponding COS domains

The working group, including Patient Research Partners (PRPs) (NG, MdW) reviewed the PROMs to achieve consensus on how to present them by COS domains. Separate scales within a multi-scale instrument as well as summed scale scores were perceived as unique instruments and mapped by their corresponding COS domains. Measurements of HRQoL were categorized as either health status surveys or health value/preference/utility assessments. The latter were reported within the COS domain 'economic cost'.

Extraction and evaluation of the methodological study quality per measurement property per instrument

The COSMIN checklist enables a critical evaluation of the methodological quality of studies investigating measurement properties[11]. A four-point system is provided to score the methodological quality of a

study per measurement property as 'excellent', 'good', 'fair' or 'poor'.[13] Four independent reviewers worked in teams of two (PH/LK, PH/AMO, PH/YYL) to reach consensus on the COSMIN ratings. A third reviewer (CT or RC) resolved disagreements. Information on score interpretation (mean (SD) of scores, floor and ceiling effects, minimally (clinically) important difference/improvement (M(C)ID/MCII), minimal detectable change (MDC) and Patient Acceptable Symptom State (PASS)) was extracted.

Evaluation of the result of the measurement properties

The results of measurement properties per instrument were evaluated (concurrently with the rating of the study methodology) as positive (+), indeterminate (?) or negative (-) per study in accordance with the quality criteria described by the 'COSMIN & Core Outcome Measures in Effectiveness Trials (COMET) collaboration'.[17]

Level of evidence for the quality of the measurement properties of PROMs in PsA

To determine the overall level of evidence for a measurement property of an instrument, data were synthesized by combining the quality of the measurement property results, the methodological study qualities and the consistency of the findings[18,19] (**Table 1**).

Table 1 Level of evidence for the quality of a measurement property

Strong (+++)	Consistent findings of ***good measurement property*** in multiple studies of good methodological quality or in one study of excellent methodological quality.
Strong (- - -)	Consistent findings of ***poor measurement property*** in multiple studies of good methodological quality or in one study of excellent methodological quality.
Moderate (++)	Consistent findings of ***good measurement property*** in multiple studies of fair methodological quality or in one study of good methodological quality.
Moderate (- -)	Consistent findings of ***poor measurement property*** in multiple studies of fair methodological quality or in one study of good methodological quality.
Limited (+)	One study of fair methodological quality with findings of ***good measurement property.***
Limited (-)	One study of fair methodological quality with findings of ***poor measurement property***
Conflicting (±)	Conflicting findings on the measurement property quality results across studies.
Unknown (?)	Only studies of poor methodological quality were identified.

Reporting the results of the evidence synthesis

As described by OMERACT[9], the COSMIN & COMET collaboration[17] and the Food And Drug

Administration (FDA)[20] guidelines, evidence on validity (especially content validity) and reliability should be prerequisites for an instrument to be considered for further evaluation/application. If an instrument does not measure what it intends to or produces unreliable estimates, it is irrelevant to test for e.g., responsiveness. Thus, in the result section of this systematic review, we have chosen to highlight the '*candidate'* instruments per COS domain that have at least limited evidence on reliability and validity and no evidence for any poor measurement properties.

The main evidence synthesis includes all studies of a PROM but conflicting evidence on measurement properties across language versions is described for '*candidate'* PROMs. Available values for Cronbach-α, interclass correlation coefficients (ICC) and floor/ceiling effects are described in the text while remaining results on measurement properties and score interpretation can be obtained from the tables.

RESULTS

Study selection

As illustrated in **Figure 1**; from 5844 unique references identified, 334 studies were eligible for further assessment. Of these, 77 reviews were excluded, as were 87 abstracts/conference papers without full-text. An additional 11 papers were added from experts and reference lists resulting in 181 studies for full-text reading. Eighty of these failed the inclusion criteria due to reasons depicted in Figure 1. Of the remaining 101 studies, clinician-reported (n=18) and composite (n=28) measures were excluded due to the focus on PROMs only, leaving 55 studies for final inclusion.

Study characteristics

The included studies were published between 1992 and 2016 and were mainly observational cohorts of PsA patients in their 4th and 5th decades of life. Most studies were performed in English speaking countries and evaluated more than one PROM (Table 2).

Characteristics of the PROMs

A total of 44 instruments covering 89 separate PROMs were evaluated (supplementary Tables B1, B2). Each PROM was mapped to the corresponding COS domain. The content, scoring and feasibility aspects of each PROM are described in supplementary Table B2.

Rating of the methodological quality and measurement property results of each study

The methodological quality ratings and ratings of the measurement property results are presented for each PROM in supplementary Table C. A further description of the rating rationale and values for score interpretation are listed per PROM in supplementary Table D.

Table 2 Characteristics of the studies

N	Sources (55 in total)	PROM(s)	N[a]	PsA(%)	Age,mean(SD)	Women(%)	Language	Country	Setting
1	Duffy (1992)[21]	AIMS1	145	100	48(13)	43	English	Canada	OPC
2	Blackmore (1995)[22]	HAQ-DI, HAQ-S, VAS stiffness$_{(HAQ)}$, VAS pain$_{(HAQ)}$	114	100	49(13)	39	English	Canada	OPC
3	Husted (1995)[23]	HAQ-SK	118	100	49(13)	39	English	Canada	OPC
4	Husted (1996)[24]	AIMS2	124	100	48(13)	40	English	Canada	OPC
5	Husted (1996)[25]	AIMS1, AIMS2	65	100	46(12)	42	English	Canada	OPC
6	Husted (1997)[26]	SF-36	113	100	51(13)	38	English	Canada	OPC
7	Taccari (1998)[27]	HAQ-DI, AIMS1	72	100	55(13)	31	Italian[b]	Italy[b]	OPC
8	Husted (1998)[28]	AIMS2, HAQ-DI, VAS pain$_{(HAQ)}$, SF-36	70	100	46(11)	39	English	Canada	OPC
9	Navsarikar (1999)[29]	DASH	50	100	49(12)	44	English	Canada	OPC
10	McKenna (2004)[30]	PsAQoL	286	100	50(13)	68	English	UK	OPC
11	Taylor (2004)[31]	BASDAI	133	100	46(19)/52(25)[c]	41/53[c]	English	New Zealand	OPC
12	Chandran (2007)[32]	FACIT-Fatigue	135	100	52(13)	41	English	Canada	OPC
13	Taylor (2007)[33]	HAQ-DI, SF-36 PF	276	49	52(14)[d]	43[d]	English	New Zealand	OPC
14	Leung (2008)[34]	HAQ-DI, BASFI, DFI, SF-36 PF	108	100	49(13)	52	Chinese	China	OPC
15	Healy (2008)[35]	PsAQoL	28	100	47(11)	50	English	UK	OPC
16	Dominguez(2009)[36]	PASE	190	19	NS	NS	English	USA	OPC
17	F.-Sueiro (2010)[37]	BASDAI	203	49	55(13)[d]	36[d]	Spanish	Spain	OPC
18	Minnock (2010)[38]	NRS Fatigue	41	100	45(13)	54	English	Ireland[b]	OPC
19	Eder (2010)[39]	BASDAI	201	100	53(14)	37	English	Canada	OPC
20	Leung (2010)[40]	SF-36, MCS, PCS	168	100	48(12)	46	Chinese	China	OPC
21	Billing (2010)[41]	PsAQoL	123	100	51(15)	53	Swedish	Sweden	OPC
22	Brodszky (2010)[42]	PsAQoL, HAQ-DI, EQ-5D-3L	183	100	50(13)	57	Hungarian	Hungary	OPC
23	Kwok (2010)[43]	VAS-pain/sleep/global/ fatigue, HAQ-DI	200	100	51(14)	59	English	Canada	OPC
24	El Miedany (2010)[44]	MultiP scales (NRS pain, NRS global (joints), NRS fatigue, mRAI, PR-TJC, NRS stiffness,	462	26.6	60(10)	72	English	UK, Egypt	OPC

		CIAQ-QoL, CIAQ-FI)							
25	Kvamme (2010)[45]	EQ-5D-3L, VAS-global/pain, mHAQ, SF-6D	4225	20.1	48(12)[d]	47[c]	Norwegian	Norway	OPC
26	Hu (2010)[46]	WTP	59	100	Range: 23-89	44	English	USA	OPC
27	Adams (2010)[47]	EQ-5D-3L, SF-6D	504	32	45(13)	52	English	Ireland	OPC
28	Adams (2011)[48]	EQ-5D-3L	504	32	45(13)	52	English	Ireland	OPC
29	Cauli (2011)[49]	VAS-global/skin/joints	319	100	52(13)	42	Multiple	Several	OPC
30	Leung (2011)[50]	SF-36, VAS pain, VAS global, HAQ-DI	20	100	48(13)/52(11)[e]	46/37[e]	Chinese	China	OPC
31	Mease (2011)[51]	HAQ-DI	161	100	47(11)	52	English	USA	RCT
32	Davis (2011)[52]	SRPQ	109	60	53(11)	37	English	Canada	OPC
33	Leung (2012)[53]	NRS-global	125	100	48(12)	48	Chinese	China	OPC
34	Leung (2013)[54]	EQ-5D-3L, SF-6D	86	100	49(13)	52	Eng/Chin	Singapore	OPC
35	Wink (2013)[55]	PsAQoL	183	100	55(13)	45	Dutch	Netherlands	OPC
36	Coaccioli (2014)[56]	PAIP	123	66	50 (22-82)	53	Italian	Italy	OPC
37	Osterhaus (2014)[57]	WPS	409	100	48(11)	55	Multiple	Several	RCT
38	Gossec (2014)[58]	PsAID-9, PsAID-12	474	100	50(13)	50	Multiple	Several	OPC
39	Torre-Al.(2014)[59]	VITACORA-19	323	65	50(19)[d]	43[d]	Spanish	Spain	OPC
40	Katchamart(2014)[60]	HAQ-DI	47	100	49(10)	55	Thai	Thailand	OPC
41	Lebwohl(2014)[61]	PSD	29/16[g]	34/50[g]	39(22-59)[f]	31[f]	English	USA	OPC
42	Chiricozzi (2015)[62]	PsoDisk	31	61.3	52(14)[f]	42[f]	Italian	Italy	OPC
43	Lubrano (2015)[63]	VAS-global	124	100	52(42-61)	53	Italian	Italy	OPC
44	Talli (2015)[64]	NRS-global/joints/skin	223	100	51(13)	51	Multiple	Several	OPC
45	Leeb (2015)[65]	SASPA	152	100	54(26-80)	46	German	Austria	OPC
46	Naegeli (2015)[66]	Worst Itch NRS	34	65	54(14)	50	English	USA	OPC
47	Wilson (2015)[67]	PSI	154	100	52(11)	63	English	USA/Canada	RCT
48	de Wit (2015)[68]	PsAID	474	100	50(13)	50	Multiple	Several	OPC
49	Tander (2016)[69]	VITACORA-19	61	100	47(12)	64	Turkish	Turkey	OPC
50	Piaserico (2016)[70]	PASE	298	19-28	NS	44[f]	Italian	Italy	OPC
51	Leung (2016)[71]	PsAQoL	98	100	52(14)	49	Eng/Chin	Singapore	OPC
52	Salaffi (2016)[72]	$PsAID_{touch}$	159	100	55(12)	61	Italian	Italy	OPC
53	di Carlo (2016)[73]	PsAID	144	100	51(13)	44	Italian	Italy	OPC

54	Cohen (2016)[74]	IPBOD	16	50	56(17)	69	English	USA	OPC
55	Cooper (2016)[75]	EQ-5D-3L	255	15	49(14)	62	Sweden	Swedish	OPC

a, Number of patients (n) often differs across the analyses within a study N in this table refers to the highest number of participants included; **b,**Presumed, not clearly stated; **c**, Axial PsA/Peripheral PsA; **d,** For the PsA group; **e**, Patient treated with TNFI <12 weeks/patients treated >12 weeks; **f**, Reported for all patients (not only PsA); **g**, Patients in the "concept elicitation"/"cognitive interview" investigation. Abbreviations: AIMS, Arthritis Impact Measurement Scale; BASDAI, Bath Ankylosing Spondylitis Activity Index; BASFI, Bath Ankylosing Spondylitis Functional index; Chin, Chinese; CIAQ-FI, Combined Inflammatory Arthritis – Functional Impairment questionnaire; CIAQ-QoL, Combined Inflammatory Arthritis – quality of life questionnaire; DASH, Disabilities of the Arm, Shoulder and Hand Outcome Measure; DFI; Dougados Functional Index; EQ-5D-3L, EuroQoL 5 Dimensions questionnaire with 3 response levels; Eng, English; FACIT-Fatigue, Functional Assessment of Chronic Illness Therapy-Fatigue Scale; Fi, Functional Index; HAQ, Health Assessment Questionnaire (HAQ-S: Spondyloarthropathy, HAQ-SK: Skin, HAQ-DI: Disability Index); IPBOD, Inverse Psoriasis Burden of Disease questionnaire; mRAI, Modified Rheumatology Attitude Index; MultiP, Multidimensional Patient Reported Outcome Questionnaire; NRS, Numeric Rating Scale; NS, Not stated; OPC, Outpatient Clinic; PAIP, Psoriatic Arthritis Impact Profile; PASE, PsA Screening and Evaluation Questionnaire; PsoDisk, abbreviation not further explained; PR-TJC, Patient-reported-tender-joint-count; PsA, Psoriatic Arthritis; PsAID, Psoriatic Arthritis Impact of Disease questionnaire; PsAQoL, PsA Quality of Life instrument; RCT, Randomised controlled trial; SASPA, Stockerau Activity Score for Psoriatic Arthritis; SF-6D, utility tool derived from SF-36 comprising six multi-level dimensions; SF-36, Medical Outcome Survey Short Form 36-item Health Survey (SF-36 MCS: Mental Component Summary, PCS: Physical Component Summary, PF: SF-36 physical function subscale; PSI, Psoriasis Symptom Inventory; SRPQ, Social Role Participation Questionnaire; VAS, Visual Analogue Scale; VITACORA-19, Spanish acronym, full name not available; WTP, Willingness to Pay Questionnaire; WPS, Work Productivity Survey.

Level of evidence on the measurement properties for each of the evaluated PROMs

Table 3 presents the *overall* evidence synthesis. Generally, most studies were of poor or fair quality resulting in limited or unknown evidence for the evaluated measurement properties. According to the results of the COSMIN analyses (supplementary Table D), frequent methodological limitations were small sample sizes, lack of information on handling of missing data, lack of information on unidimensionality when assessing internal consistency, insufficient methods for examining/reporting content validity, inappropriate statistical methods for testing responsiveness, and lack of hypotheses and psychometric information on comparators when testing construct validity.

Evidence for PROMS measuring PsA core domains

MUSCULOSKELETAL DISEASE ACTIVITY.

The core domain of musculoskeletal disease activity is currently measured using a combination of physician assessments (clinical examination) and PROMs, and depending on the purpose of the study also biologic inflammatory markers and/or assessments of PsA pathophysiology using tissue imaging techniques. Six PROMs that aim to evaluate the concept of patient reported disease activity were retrieved (Table 3). The Stockerau Activity Score for Psoriatic Arthritis (SASPA) in German was currently the best candidate based on limited evidence for unidimensionality, internal consistency (Cronbach-α=0.875) as well as structural validity by factor analysis (supplementary Table C and D). SASPA is short, free and easy to score (supplementary Table B2). The main limitations of SASPA are the unknown content validity and only the original German version was evaluated. SASPA is available in English but without information on the quality of the translation or cross-cultural validation.

SKIN DISEASE ACTIVITY

Three instruments were found that aim to measure patient reported skin disease activity (Table 3). Strong evidence for content validity of the Psoriasis Symptom Diary (PSD) was obtained while information on remaining measurement properties was not available in PsA. Based on results from Rasch and principal

component analysis, the Psoriasis Symptom Inventory (PSI) appeared the best available PROM having moderate evidence for unidimensionality, internal consistency (Cronbach α=0.95) and structural validity, and limited evidence for responsiveness, test-retest reliability (ICC=0.70) and construct validity (external relationships and known group validity). The main limitations of PSI include item floor effects (up to 37% at baseline) (supplementary Table D).

PAIN

Six PROMs were evaluated (Table 3). None of these had evidence on both reliability and validity. The Medical Outcome Survey Short Form 36-item Health Survey Bodily Pain subscale (SF-36 BP) was evaluated by Chinese and English studies generating moderate and limited evidence for construct validity regarding internal and external relationships, respectively. Evidence for unidimensionality of the BP scale was not provided by the studies reporting on Cronbach-α (0.80-0.91) leading to no overall evidence for internal consistency. Information on floor effects (1.2%), ceiling effects (3.0%) and MID was provided (supplementary Table D). The main limitations of SF-36 BP are the unknown evidence for reliability and content validity, and the requirement of software to calculate scores (supplementary Table B2). The visual analogue scale (VAS) of pain (1 week recall time) had limited evidence for construct validity (external relationships) (Table 3), and MID was reported (Table 3, and supplementary Table C and D).

PATIENT GLOBAL

Eight measures of Patient Global (PtG) were identified and included VAS and numeric rating scales (NRS) with varying recall periods. The phrasing of the PtG item addressed the impact on overall well-being of either 1) arthritis, 2) psoriasis, or 3) PsA (as a whole) as described in supplementary Table B2. Only the VAS of PtG due to PsA (1 week recall) had evidence of both validity and reliability in PsA including limited evidence for construct validity (external relationships) and moderate evidence for test-retest reliability (ICC (95%CI) =0.87(0.83-0.90)). Values of MID, PASS and MCII were reported across languages and recall versions of VAS PtG (Table 3, supplementary Tables C-E). The NRS of PtG due to PsA (1 week recall) had

moderate evidence for construct validity (external relationships and known group validity) and floor/ceiling effects were reported up to ~ 8 %/3 % (Table 3 and supplementary Table D).

PHYSICAL FUNCTION

Twenty-three PROMs were evaluated (Table 3), and three of these had evidence on both reliability and validity including the Bath Ankylosing Functional Index (BASFI), the SF-36 Physical Function subscale (SF-36 PF) and the Health Assessment Questionnaire Disability Index (HAQ-DI). Based on evidence from English and Chinese studies using Rasch analysis and principal component analysis, the SF-36 PF was the best candidate with strong evidence for unidimensionality, internal consistency (Cronbach α=0.91-0.92) and good structural validity. Evidence for construct validity was moderate and limited for internal and external relationships, respectively (Table 3). Floor and ceiling effects were less than 10% and MID was reported (supplementary Table D). The HAQ-DI was the most frequently assessed instrument for this domain and had strong evidence for good internal consistency and structural validity (Table 3). However Rasch analysis suggested better properties for the SF-36 PF in a study that compared the two instruments.[33] HAQ-DI was limited by floor effect (up to 50%) and had conflicting evidence on construct validity across languages (supplementary Tables C-E).

HEALTH RELATED QUALITY OF LIFE/LIFE IMPACT

Ten PROMs were identified (Table 3). Of these, the Psoriatic Arthritis Impact of Disease (PsAID) questionnaire, the PsA Quality of Life instrument (PsAQoL) and the VITACORA-19 (Spanish and Italian versions) all had some evidence on both reliability and validity. PsAID was translated and evaluated in several languages during the development phase and appeared a good candidate based on strong evidence for content validity and moderate evidence for good test-retest reliability and for good construct validity (external relationships) of the 12-item version (PsAID-12). Similar findings existed for PsAID-9 except that evidence for construct validity was limited. Floor/ceiling effects of PsAID were <1%, and values for PASS were provided (supplementary Table D). The PsAQoL was assessed in several language versions (supplementary Tables C-E) generating strong evidence for unidimensionality and internal consistency

(Cronbach α=0.91) and moderate evidence for test-retest reliability and structural, construct validity (external relationships and known group validity) (Tables 3). Moderate and strong evidence for content validity was available for the English and Swedish versions of PsAQoL, while limited evidence for poor content validity was achieved by a Dutch study where approximately half of the patients suggested a lack of items, resulting in overall conflicting evidence for this property (supplementary Tables C-E). Floor effect of PsAQoL was up to 19% (supplementary Table D). VITACORA-19 was evaluated in Spanish (origin) and in Turkish resulting in moderate evidence for test-retest reliability (ICC=0.94), content validity and construct validity (external relationships) as well as limited evidence for unidimensionality, internal consistency (Cronbach α= 0.95) and good structural validity. Floor/ceiling effects were <1% and MCID was defined (supplementary Table D). No formal English translation or cross-cultural validation was available.

FATIGUE

Four instruments were identified (Table 3). Evidence for validity and reliability was only available for the Functional Assessment of Chronic Illness Therapy-Fatigue scale (FACIT-Fatigue) including limited evidence for good test-retest reliability (ICC=0.95) and construct validity (external relationships) (Table 3, supplementary Table D).

PROMs measuring domains of the middle circle of the PsA COS

PARTICIPATION

Eleven PROMs were evaluated (Table 3). The three subscales of the Social Role Participation Questionnaire were the only measurements with evidence of both reliability and validity including limited evidence for good test-retest reliability, content validity and construct (external relationships and known group) validity. The Work Productivity Survey had limited evidence for good construct validity and responsiveness but high floor effects found for certain items (73.7% (item 2) and 77.3% (item 8)) (Table 3, supplementary Table D). The SF-36 role emotional, role physical and social functioning subscales had moderate evidence for good construct validity (hypotheses testing regarding known groups, internal and external relationships).

EMOTIONAL WELL-BEING

Nine instruments were identified from Chinese and English studies but none had evidence on both validity and reliability (Table 3). The most information was available for the SF-36 Mental Health subscale (SF-36 MH) and the SF-36 mental component summary (MCS) including moderate evidence for good construct (internal relationships) and structural validity, respectively (Table 3, supplementary Table D).

ECONOMIC COST

Four instruments were available (Table 3) but none of these had evidence for both reliability and validity. Evidence for construct validity (external relationships) was available for the EuroQol-5 Domain 3 level (EQ 5D-3L) (moderate) and the SF-6D (derived from SF-36) and Willingness-to-pay questionnaire (both limited). Differences in utility estimates from EQ-5D versus SF-6D, score distribution, floor/ceiling effects, PASS and MCII information were reported (supplementary Table D).

PROMs measuring domains of the COS research agenda (outer circle)

SLEEP

One study assessed VAS Sleep providing information on score interpretation (Table 3, supplementary Table D).

STIFFNESS

Two measurements, VAS Stiffness and the NRS Stiffness were evaluated (Table 3) but the evidence for measurement properties remained unknown (Table 3, supplementary Table D).

PROMS measuring domains not included in the COS

SF-36 general health subscale (GH) and the Arthritis Impact Measurement (AIMS 2) Social Support scale were evaluated but evidence for measurement properties was not achieved (Tables 3, supplementary Table D).

Table 3 Level of evidence for measurement properties per PROM listed by matching COS domain

PROMs by COS Domains (n=89)	Reliability COSMIN BOX (A-C)			Validity COSMIN BOX (D-H)					Responsiveness COSMIN BOX (I)	Info on score interpretation (values are provided in suppl. Table D)
	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypotheses testing	Cross-cult. Validity	Criterion validity	Sensitivity to change	
	A	B	C	D	E	F	G	H	I	
MSK DISEASE ACTIVITY, patient reported aspects (n=6)										
BASDAI[31,37,39]	?					±			?	F/C
SASPA[65]	+				+	?			?	
PASE-total[36,70]		?				+	*A*		+	
PASE-symptom[36,70]		?				+	*A*		+	
PASE-function[36,70]		?				+	*A*		+	
PR-TJC[44]				?		?				
SKIN DISEASE ACTIVITY, patient reported aspects (n=3)										
PSI[67]	++	+			++	+			+	F/C
PSD[61]				+++						
Worst itch NRS[66]				+						
PAIN (n=6)										
VAS Pain (1 week recall)[22,28,43,50]						+			?	MID
VAS Pain (recall NS)[45]										MCII, PASS
NRS Pain (1 week recall)[44]		?		?		?				
SF-36 BP[26,28,40,50]	?					+/++ *b*			?	MID, F/C
AIMS1 Pain[21,25,27]						++			?	
AIMS2 Pain[24,25,28]	?					+			?	
PATIENT GLOBAL (n=8)										
<u>Patient global due to psoriasis</u>										
NRS (1 week recall)[64]						+				F/C
VAS (1 week recall)[49]		++				?				
<u>Patient global due to arthritis</u>										
NRS (1 week recall)[64]						+				F/C
NRS (1 day recall)[44]				?		?				
VAS (1 week recall)[49]		++				?				
<u>Patient global due to PsA</u>										
NRS (1 week recall)[53,64]						++				F/C

VAS (1 week recall)[43,49,50,63]		++				+			?	MID
VAS (recall NS)[45]										MID, PASS, MCII

Table 3 cont.	**Reliability COSMIN BOX (A-C)**			**Validity COSMIN BOX (D-H)**					**Responsiveness COSMIN BOX (I)**	**Info on score interpretation (values are provided in suppl. Table D)**
PROMs by COS Domains	**Internal consistency**	**Reliability**	**Measurement error**	**Content validity**	**Structural validity**	**Hypotheses testing**	**Cross-cult. Validity**	**Criterion validity**	**Sensitivity to change**	
	A	B	C	D	E	F	G	H	I	
PHYSICAL FUNCTION (n=23)										**Interpretability**
DFI[34]	– –				– –	?				F/C
DASH[29]						– –				
BASFI[34]	++				++	?				F/C
HAQ-DI [22,27,28,33,34,42,43,50,51,60]	+++				+++	±			?	F/C, MID
HAQ-S[22]						–				
HAQ-SK[23]						?				
mHAQ[45]										PASS, MCII
SF-36 PF[26,28,33,34,40,50]	+++				+++	+/++*b*			?	F/C, MID
SF-36 PCS[40,50]					++	?			?	
MultiP CASQ-FI[44]	?				?	?				
AIMS1 Mobility[21]						-				
AIMS1 Physical[21,27]						±				
AIMS1 Dexterity[21]						+				
AIMS1 House[21]						+				
AIMS1 ADL[21]						–				
AIMS1 PC[25]									?	
AIMS2 PC[25,28]									?	
AIMS2 Mobility[24]						+				
AIMS2 Physical[24]						+				
AIMS2 Dexterity[24]						+				
AIMS2 Selfcare[24]						–				
AIMS2 House[24]						–				
AIMS2 Arm F.[24]						+				

Table 3 cont. **PROMs by COS Domains**	**Reliability COSMIN BOX (A-C)**			**Validity COSMIN BOX (D-H)**					**Responsiveness COSMIN BOX (I)**	**Info on score interpretation (values are provided in suppl. Table D)**
	Internal consistency	**Reliability**	**Measurement error**	**Content validity**	**Structural validity**	**Hypotheses testing**	**Cross-cult. Validity**	**Criterion validity**	**Sensitivity to change**	
	A	B	C	D	E	F	G	H	I	
HRQoL/LIFE IMPACT (n=10)										
PsAQoL[30,35,41,42,55,71]	+++	++	?	±	++	++	*a*		?	F/C
AIMS1 Global[27]						?				
PsAID-9[58,68]	*c*	++		+++		+	*a*		?	PASS, F/C
PsAID-12[58,68,73]	*c*	++		+++	*c*	++	*a*		?	PASS, F/C
$_{touch}$PsAID-12[72]						+		+*d*		MDA cut-off
PAIP[56]						?				
VITACORA-19[59,69]	+	++		++	+	++	*a*		?	MCID, F/C
PsoDisk[62]									?	
MultiP CIAQ-QoL[44]		?		?		?				
IBOD[74]	*c*			?		?				
FATIGUE (n=4)										
FACIT-Fatigue[32]	?	+				+				
NRS fatigue[38,44]		?		?		?			?	
VAS fatigue[43]										MID
SF-36 VT[26,40,50]	?					–/++*b*				MID, F/C
PARTICIPATION (n=11)										
SRPQ-IM[52]	?	+	?	+		+				MDC
SRPQ-ST[52]	?	+	?	+		+				MDC
SRPQ-SR[52]	?	+	?	+		+				MDC
WPS[57]						+			+	F/C
AIMS1 SA[21]						?				
AIMS2 SA[24]						?				
AIMS2 Work[24]						?				
AIMS2 SC[28]									?	
SF-36 RE[26,40,50]	?					?/++ *b*			?	
SF-36 RP[26,40,50]	?					–/++ *b*			?	

SF-36 SF[26,28,40,50]	?					?/++ *b*			?	

Table 3 cont **PROMs by COS Domains**	**Reliability COSMIN BOX (A-C)**			**Validity COSMIN BOX (D-H)**					**Responsiveness COSMIN BOX (I)**	**Info on score interpretation (values are provided in suppl. Table D)**
	Internal consistency	**Reliability**	**Measurement error**	**Content validity**	**Structural validity**	**Hypotheses testing**	**Cross-cult. Validity**	**Criterion validity**	**Sensitivity to change**	
	A	B	C	D	E	F	G	H	I	
EMOTIONAL WELL-BEING (n=9)										
SF-36 MH[26,28,40,50]	?					++ *b*			?	MID
SF-36 MCS[40,50]					++	?			?	
MultiP mRAI[44]		?		?		?				
AIMS1 Psyc.C.[25]									?	
AIMS1 Anxiety[21]						?				
AIMS1 Depression[21]						?				
AIMS2 Mood[21]						?				
AIMS2 Tension[21]						?				
AIMS2 Psyc.C.[25,28]									?	
ECONOMIC COST (n=4)										
EQ-5D [42,45,47,48,54,75]						++			?	MCII, PASS, F/C
EQ-5D-revised[48]						?			?	Score distribution
SF-6D[45,47,54]						+			?	PASS, MCII, F/C
WTP[46]				?		+				
SLEEP (n=1)										
VAS sleep[43]										MID
STIFFNESS (n=2)										
NRS stiffness[44]				?		?				
VAS stiffness[22]						?				
NON-COS Domains (n=2)										
SF-36 GH[26,40,50]	?					– /– – *b*				
AIMS2 Social Support[24]						?				

Empty cells reflect that the measurement property was not evaluated by any study for the given instrument**. T**able 2 explains the grading of evidence (+/-/?).

aOnly translation, no cross-cultural validation. According to COSMIN, only studies that address measurement invariance (e.g. multiple group factor analyses or DIF) between countries (or other groups) are considered real cross-cultural validity studies. ***b***Construct validity – hypotheses testing was assessed regarding the internal relationships

(scale assumptions) and not relation to external measurements. [c] Questionnaire seems to be based on a formative model why scoring of internal consistency and structural validity is not relevant. [d] PsAID touch version was compared to paper version which was considered as gold standard. Abbreviations: AIMS, Arthritis Impact Measurement Scales (ADL, Activity of daily living; Arm F., Arm Function; House, Household; PC, Physical component score; Psyc.C., Psychological component score; SA, Social Activity, SC, Social component score); BASDAI, Bath Ankylosing Spondylitis Activity Index; BASFI, Bath Ankylosing Spondylitis Functional index; CIAQ-FI, Combined Inflammatory Arthritis – Functional Impairment questionnaire; CIAQ-QoI, Combined Inflammatory Arthritis – quality of life questionnaire; COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; DASH, Disabilities of the Arm, Shoulder and Hand Outcome Measure; DFI, Dougados Functional Index; EQ-5D-3L, EuroQoL 5 Dimensions questionnaire with 3 response levels; FACIT-Fatigue, Functional Assessment of Chronic Illness Therapy-Fatigue scale; F/C, Floor/Ceiling effect; HAQ, Health Assessment Questionnaire (HAQ-S: Spondyloarthropathy, HAQ-SK: Skin, HAQ-DI: Disability Index); IPBOD, Inverse Psoriasis Burden of Disease questionnaire; MCID, Minimal clinically important difference; MDA, Minimal disease activity; MDC, minimal detectable change; MCII, Minimal clinical important improvement; MIC, Minimal important change; MID, Minimal important difference; mRAI, Modified Rheumatology Attitude Index; MultiP, Multidimensional Patient Reported Outcome Questionnaire; NRS, Numeric Rating Scale; NS, Not stated; PAIP, Psoriatic Arthritis Impact Profile; PASE, PsA Screening and Evaluation Questionnaire; PASS, Patient acceptable symptom state; PGA, Patient Global Assessment; PR-TJC, Patient-reported-tender-joint-count; PsAID, Psoriatic Arthritis Impact of Disease questionnaire; PsAQoL, PsA Quality of Life instrument; PSD; Psoriasis symptom diary; PSI, Psoriasis Symptom Inventory; Psodisk questionnaire, no full spelling available; SASPA, Stockerau Activity Score for Psoriatic Arthritis; SF-6D, utility tool derived from SF-36 comprising six multi-level dimensions; SF-36, Medical Outcome Survey Short Form 36-item Health Survey (SF-36 subscales: BP, Bodily Pain; GH, General Health; MCS, Mental Component Summary; MH, Mental Health; PCS, Physical Component Summary, PF, physical function; RE, Role Emotional; RP, Role Physical; SF, Social Functioning; VT, Vitality); SRPQ, Social Role Participation Questionnaire; VAS, Visual Analogue Scale; VITACORA-19, Spanish acronym, full name not available; WTP, Willingness to pay questionnaire; WPS, Work Productivity Survey.

DISCUSSION

Core outcome measurement sets (COMS) aim to ensure the best possible evaluation of the domains in a core outcome set (COS) for a specific disease, providing comparability across study results and enhancement of evidence-based health care decisions. While previous studies have provided overviews of commonly used instruments in PsA,[76,77] this review provides a systematic identification, characterization and evidence synthesis of measurement properties of all PROMs evaluated in PsA, which constitutes an important step in the GRAPPA-OMERACT process of developing a PsA COMS.

PROMs with at least some evidence on both reliability and validity are available for six of the eight mandatory ("inner circle") COS domains including MSK disease activity (SASPA), skin disease activity (PSI), patient global (VAS global), physical function (SF-36 PF, HAQ-DI, BASFI), HRQoL/life impact (PsAID-9, PsAID-12, PsAQoL, VITACORA-19) and fatigue (FACIT-Fatigue).

Instruments with *strong* evidence for any measurement property included HAQ-DI and SF-36 PF (physical function domain), PSD (skin disease activity domain), PsAID-9, PsAID-12 and the English version of PsAQoL (HRQoL/life impact domain). The PSD, PsAID-9, PsAID-12, and English PsAQoL had strong evidence on content validity, a property that was sparsely investigated for most other PROMs. Content validity is considered a prerequisite for applicability of PROMS in PsA clinical trials as emphasized by the FDA, OMERACT and the COSMIN-COMET initiative.[17,20,78] Thus, unknown content validity of PROMS is a serious shortcoming that needs attention in PsA – as well as in other rheumatic diseases.[58,79,80]

No PROM with evidence on both reliability and validity was available for the mandatory COS domains of systemic inflammation and pain. The absence of a good PROM for assessment of pain is especially critical as clinicians and patients have considered this patient-reported domain extremely important according to former studies.[5,58] Future research should gain more information on the measurement properties of the SF-36 pain subscale, VAS pain and the AIMS pain scale that all had some evidence of validity in PsA according to this SLR.

Furthermore, data from the PsAID study could provide additional evidence for use of the individual NRS for several of the COS domains, including pain. The applicability of the Patient Reported Outcomes Measurement Information System (PROMIS) for measuring pain as well as other domains of the PsA COS may also be considered.[81] PROMIS provides multiple unidimensional instruments that can be administered as fixed short forms as well as computer adaptive tests. The SF-36 subscales assess three inner core domains (pain, physical function and fatigue/vitality) and a visual representation of the multiple life impact/HRQoL domains can be generated through spydergrams.[82] It may seem practical to use a questionnaire with multiple scales that cover several domains in one application. However, it is more important to endorse the best instrument per domain and further research must be done on the measurement properties of SF-36 subscales in PsA.

All language versions of a PROM were lumped in the main evidence synthesis of this review to achieve as much information as possible per instrument. This strategy underscores the importance of collecting sufficient evidence on cross-cultural validity prior to international application of a PROM. For instance, the German SASPA (MSK disease activity) and the Italian/Turkish VITACORA-19 (HRQoL) both have some evidence for reliability and validity but translation (and cross-cultural validation) into the most common languages (English at least) is warranted. Furthermore, the evidence for content validity of PsAQoL and construct validity of HAQ-DI was rated as conflicting in the overall synthesis mainly due to diverging results across language versions. Given the limited number and quality of the included studies, future studies of high methodological standards should clarify if such differences truly exist and if they are cross-culturally related. Several studies evaluated the measurement properties of a translated questionnaire but according to COSMIN, only studies that address measurement invariance (e.g. multiple group factor analyses or DIF) between countries (or other groups) are considered real cross-cultural validity studies.

Few studies with sufficient methodology for assessing responsiveness were identified. Although reliability and validity were considered preconditions for potential PROMs, the COMS is being developed for clinical

trials for which measuring the true amount of change in a construct during an intervention is often the primary goal. Therefore, responsiveness of promising instruments needs to be clarified in future studies. The evidence for measurement properties of PROMs measuring skin disease activity was limited since we included only studies with at least 50% of the population comprising PsA patients (or PsA subgroup results). This strategy may be conservative, for instance additional information on the candidate instrument PSI as well as on PSD would have been achieved by including studies of psoriasis.[83-86] Nevertheless, our strategy ensures that the evidence obtained applies to patients with PsA as a whole. Strengths of this GRAPPA-OMERACT study constitute the international collaboration including experts in PsA, measurement and systematic review technique as well as patient research partners. Adherence to the COSMIN guidelines guaranties homogeneity and transparency in the assessment of methodology and rating of measurement properties across studies. Study limitations include, as for reviews in general, that negative findings might have been underreported due to publication bias. Selection bias due to exclusion of non-English full-text papers may have led to underreporting of the (cross-cultural) evidence for some instruments. However we believe this was minimized as only five studies were excluded for this reason. This review did not include RCTs or longitudinal observational studies that only provide indirect evidence for measurement properties of instruments used for assessing the outcomes of interest. We acknowledge that great amounts of indirect evidence are available and valuable in the COMS development. However the identification, selection and evaluation strategies needed for such studies do not comply with the methodology of the current review. Further analyses are currently underway by parallel work streams evaluating the data from PROMs collected in recently conducted RCTs of interventional therapies in PsA to fully adhere to the OMERACT procedure of COMS development.

This study provides an evidence based overview of measurement properties of PROMs per COS domain. We have highlighted the current knowledge gaps, and provided an overview of available data on score interpretation, feasibility and content for each PROM. This constitutes a relevant starting point for stakeholders to decide on the overall applicability of the PROMs, and provides opportunities to improve

existing data by targeted research strategies.[6,10] This is indeed warranted as several of the PROMs with elusive measurement properties are widely used in PsA trials and clinics today. [77] Some COS domains may be more appropriately assessed by non-PROM instruments such as biomarkers and clinical assessments, and parallel work streams within GRAPPA-OMERACT are collecting psychometric evidence for the use of such tools in PsA. These research initiatives will in addition to the psychometric evidence for PsA PROMs presented in this review inform the consecutive stages of developing a COMS for PsA.

CONTRIBUTORS: All of the authors fulfil the following criteria: substantial contributions to the conception or design of the work; or the acquisition, analysis or interpretation of data for the work; and drafting the work or revising it critically for important intellectual content; and final approval of the version to be published; and agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGEMENT: The study was financially supported by the Parker Institute, Copenhagen University Hospital, Bispebjerg and Frederiksberg, DK, Danish Rheumatism Association (R124-A3278-B984), and Department of Rheumatology, Rigshospitalet, Gentofte Hospital, DK. The Parker Institute, Bispebjerg and Frederiksberg Hospital is supported by a core grant from the Oak Foundation (OCAY-13-309). None of the funding sources had any influence on the study design, on data collection, data synthesis, data interpretation, writing the report, or the decision to submit the manuscript for publication.

Supplementary material:

Supplementary Table A: Search strategy

SEARCH STRATEGY INCORPORATING THE MEASUREMENT PROPERTY FILTERS BY TERWEE et al[1]

Final search (#5) "Search (#1 OR #2 OR #3) AND #4"

PUBMED SEARCH FILTER (SEPARATED INTO 3 SECTIONS FOR CLARITY)

Search #1: ((replicab*[Title/Abstract] OR repeated[Title/Abstract]) AND (measure[Title/Abstract] OR measures[Title/Abstract] OR findings[Title/Abstract] OR result[Title/Abstract] OR results[Title/Abstract] OR test[Title/Abstract] OR tests[Title/Abstract])) OR ("meaningful change"[Title/Abstract]) OR ((small*[Title/Abstract]) AND (real[Title/Abstract] OR detectable[Title/Abstract]) AND (change[Title/Abstract] OR difference[Title/Abstract])) OR ((minimal[Title/Abstract] OR minimally[Title/Abstract] OR clinical[Title/Abstract] OR clinically[Title/Abstract]) AND (important[Title/Abstract] OR significant[Title/Abstract] OR detectable[Title/Abstract]) AND (change[Title/Abstract] OR difference[Title/Abstract]))

Search #2: (((stability[Title/Abstract]) OR interrater[Title/Abstract]) OR inter-rater[Title/Abstract]) OR intrarater[Title/Abstract]) OR intra-rater[Title/Abstract]) OR intertester[Title/Abstract]) OR inter-tester[Title/Abstract]) OR intratester[Title/Abstract]) OR intra-tester[Title/Abstract]) OR interobserver[Title/Abstract]) OR inter-observer[Title/Abstract]) OR intra-observer[Title/Abstract]) OR intraobserver[Title/Abstract]) OR intertechnician[Title/Abstract]) OR inter-technician[Title/Abstract]) OR intra-technician[Title/Abstract]) OR intratechnician[Title/Abstract]) OR interexaminer[Title/Abstract]) OR intraexaminer[Title/Abstract]) OR intra-examiner[Title/Abstract]) OR inter-examiner[Title/Abstract]) OR interassay[Title/Abstract]) OR inter-assay[Title/Abstract]) OR intra-assay[Title/Abstract]) OR intraassay[Title/Abstract]) OR interindividual[Title/Abstract]) OR inter-individual[Title/Abstract]) OR intra-individual[Title/Abstract]) OR intraindividual[Title/Abstract]) OR interparticipant[Title/Abstract]) OR intra-participant[Title/Abstract]) OR inter-participant[Title/Abstract]) OR intraparticipant[Title/Abstract]) OR kappa*[Title/Abstract]) OR repeatab*[Title/Abstract]) OR generaliza*[Title/Abstract]) OR generalisa*[Title/Abstract]) OR concordance[Title/Abstract]) OR ((intraclass[Title/Abstract] AND correlation*[Title/Abstract]))) OR (("intra-class"[Title/Abstract] AND correlation*[Title/Abstract]))) OR discriminative[Title/Abstract]) OR "known group"[Title/Abstract]) OR "factor analysis"[Title/Abstract]) OR "factor analyses"[Title/Abstract]) OR dimension*[Title/Abstract]) OR subscale*[Title/Abstract]) OR ((multitrait[Title/Abstract] AND scaling[Title/Abstract]) AND (analysis[Title/Abstract] OR analyses[Title/Abstract])))) OR "item discriminant"[Title/Abstract]) OR "inter scale correlation*"[Title/Abstract]) OR "interscale correlation*"[Title/Abstract]) OR error[Title/Abstract]) OR errors[Title/Abstract]) OR "individual variability"[Title/Abstract]) OR ((variability[Title/Abstract] AND analysis[Title/Abstract]))) OR ((variability[Title/Abstract] AND values[Title/Abstract]))) OR ((uncertainty[Title/Abstract] AND (measurement*[Title/Abstract] OR measuring[Title/Abstract])))) OR sensitiv*[Title/Abstract] OR responsive*[Title/Abstract])) OR "ceiling effect"[Title/Abstract]) OR "floor effect"[Title/Abstract]) OR "item response model"[Title/Abstract]) OR "Item Response Theory"[Title/Abstract]) OR Rasch[Title/Abstract]) OR "differential item functioning"[Title/Abstract]) OR "computer adaptive testing"[Title/Abstract]) OR "item bank"[Title/Abstract]) OR "cross-cultural equivalence"[Title/Abstract]))

Search #3: (((((((((((((((((((((((((instrumentation[MeSH Subheading]) OR methods[MeSH Subheading]) OR Validation studies[Publication Type]) OR Comparative study[Publication Type]) OR "psychometrics"[MeSH Terms]) OR psychometr*[Title/Abstract]) OR clinimetr*[Title/Abstract]) OR clinometr*[Title/Abstract]) OR "outcome assessment (health care)"[MeSH Terms]) OR "outcome assessment"[Title/Abstract]) OR "outcome measure*"[Text Word]) OR "observer variation"[Text Word] OR "health status indicators"[MeSH Terms]) OR "reproducibility of results"[MeSH Terms]) OR reproducib*[Title/Abstract]) OR "discriminant analysis"[MeSH Terms]) OR reliab*[Title/Abstract]) OR unreliab*[Title/Abstract]) OR valid*[Title/Abstract]) OR coefficient[Title/Abstract]) OR homogeneity[Title/Abstract]) OR homogeneous[Title/Abstract] OR "internal consistency"[Title/Abstract]) OR "cronbach* alpha*"[Title/Abstract] OR ((item*[Title/Abstract]) AND (selection*[Title/Abstract] OR correlation*[Title/Abstract] OR reduction*[Title/Abstract])) OR agreement[Title/Abstract]) OR precision[Title/Abstract]) OR imprecision[Title/Abstract]) OR "precise values"[Title/Abstract]) OR "test retest"[Title/Abstract]

TARGET POPULATION

Search #4: ((((((((((((((((((((psoriatic arthritis[MeSH Terms]) OR psoriatic arthrit*[Title/Abstract]) OR psoriatic arthropath*[Title/Abstract]) OR Psoriatic spondylarthropath*[Title/Abstract]) OR Psoriatic joint[Title/Abstract]) OR Psoriatic joints [Title/Abstract]) OR (Psoriasis[Title/Abstract] AND spondylarthropath*[Title/Abstract])) OR arthritis psoriatica[Title/Abstract]) OR psoriatic polyarthritis[Title/Abstract]) OR psoriatic rheumatism[Title/Abstract]) OR psoriatic spondylit*[Title/Abstract]) OR psoriatic spondylo*[Title/Abstract]) OR arthropathic psoriasis[Title/Abstract]) OR arthritis mutilans[Title/Abstract]) OR (psoriasis[Title/Abstract] AND arthritis[Title/Abstract])) OR (psoriasis[Title/Abstract] AND arthropath*[Title/Abstract])) OR (psoriatic joint*[Title/Abstract])) OR (psoriasis pustulosa arthropat*[Title/Abstract])) OR (psoriasis[Title/Abstract] AND enthes*[Title/Abstract])) OR (psoriasis[Title/Abstract] AND dactylit*[Title/Abstract])) OR (psoriasis[Title/Abstract] AND spondylit*[Title/Abstract])) OR (psoriasis[Title/Abstract] AND spondylo*[Title/Abstract])) OR (psoriasis[Title/Abstract] AND SpA[Title/Abstract])) OR (psoriasis[Title/Abstract] AND PsA[Title/Abstract]) OR (psoriasis[Title/Abstract] AND (joint[Title/Abstract] OR joints[Title/Abstract]))

EMBASE SEARCH STRATEGY

Final search (#3): Search (#1 AND #2)

MEASUREMENT PROPERTY FILTER FOR EMBASE

SEARCH #1: intermethod comparison*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword] OR inter method comparison*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword] OR data collection method*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword] OR. data collection method/ or interview/ or observational method/ or questionnaire/ OR validation study.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword] OR feasibility study.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword] OR pilot study.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword] OR psychometr*.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword] OR exp psychometry/ OR reproducib*.ti,ab. OR audit.ti,ab. OR clinometr*.ti,ab. OR clinimetr*.ti,ab. OR observer variation.ti,ab. OR exp observer variation/OR discriminant analysis.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword] OR exp discriminant validity/ or exp predictive validity/ or exp content validity/ or exp face validity/ or exp construct validity/ or exp qualitative validity/ or exp validity/ or exp external validity/ or exp consensual validity/ or exp convergent validity/ or exp concurrent validity/ or exp internal validity/ or exp criterion related validity/ OR exp reliability/ OR reliability.ti,ab. OR coefficient.ab,ti. OR internal consistency.ab,ti. OR (cronbach and alpha*).ti,ab. OR item correlation*.ti,ab. OR item selection*.ti,ab. OR item reduction*.ti,ab. OR exp diagnostic accuracy/ or exp measurement accuracy/ or exp accuracy/ or exp dimensional measurement accuracy/ or exp diagnostic test accuracy study/ OR imprecision.ti,ab. OR (test and retest).ti,ab. OR interrater.ti,ab. OR inter-rater.ti,ab.OR intra-rater.ti,ab. OR intrarater.ti,ab.OR interobserver.ti,ab. OR inter observer.ti,ab. OR intra observer.ti,ab. OR intraobserver.ti,ab. OR interexaminer.ti,ab. OR inter examiner.ti,ab. OR intra examiner.ti,ab. OR intraexaminer.ti,ab. OR interindividual.ti,ab.OR inter-individual.ti,ab.OR intraindividual.ti,ab.OR intra-individual.ti,ab.OR interparticipant.ti,ab.OR inter participant.ti,ab. OR intra participant.ti,ab. OR intraparticipant.ti,ab. OR intertechnician.ti,ab.OR inter technician.ti,ab. OR intratechnician.ti,ab. OR intra technician.ti,ab. OR (kappa and value).ti,ab. OR (kappa and statistics).ti,ab. OR (repeated and measure*).ti,ab. OR (repeated and finding*).ti,ab. OR (repeated and result*).ti,ab. OR (repeated and test*).ti,ab. OR repeatab*.ti,ab. OR (replicab* and measure*).ti,ab. OR (replicab* and finding*).ti,ab. OR (replicab* and result*).ti,ab. OR (replicab* and test*).ti,ab.
(intra-class and correlation).ti,ab. OR (intraclass and correlation).ti,ab. OR factor structure.ti,ab. OR factor analys*.ti,ab. OR dimensionality.ti,ab. OR multitrait scaling analys*.ti,ab. OR item discriminant.ti,ab. OR interscale correlation*.ti,ab. OR inter-scale correlation*.ti,ab. OR (error* and measurement).ti,ab. OR interval variability.ti,ab. OR responsiveness.ti,ab. OR minimal detectable.ti,ab. OR meaningful change.ti,ab. OR ceiling effect.ti,ab. OR floor effect.ti,ab. OR item response model.ti,ab. OR item response theory.ti,ab. OR rasch.ti,ab. OR differential item functioning.ti,ab. OR touch screen.ti,ab. OR item bank.ti,ab. OR cross-cultural equivalence.ti,ab. OR crosscultural equivalence.ti,ab.

TARGET POPULATION

Search #2 exp psoriatic arthritis/ OR arthritis psoriatica.ti,ab. OR psoriatic arthritis.ti,ab. OR psoriatic polyarthritis.ti,ab. OR psoriatic rheumatism.ti,ab. OR psoriatic spondylit*.ti,ab. OR psoriatic joint.ti,ab. OR psoriatic spondylo*.ti,ab. OR psoriatic joints.ti,ab. OR psoriatic spondylarthropath*.ti,ab. OR (psoriasis and spondylo*).ti,ab. OR psoriasis pustulosa arthropat*.ti,ab. OR (psoriasis and spondylit*).ti,ab. OR arthropathic psoriasis.ti,ab. OR psoriatic arthropath*.ti,ab. OR arthritis mutilans.ti,ab. OR (psoriatic and arthritis).ti,ab. OR (psoriasis and enthes*).ti,ab. OR (psoriasis and dactylit*).ti,ab. OR (psoriasis and spondylarthropath*).ti,ab. OR (psoriasis and SpA).ti,ab. OR (psoriasis and PsA).ti,ab. OR (psoriasis and joints).ti,ab. OR (psoriasis and joints).ti,ab. OR (psoriasis and arthropath*).ti,ab.

Abbreviations: Ti,ab.: Title, abstract. Exp: explode

PsycINFO SEARCH STRATEGY: Only the target population (as described above) was used as search criteria (no measurement property terms applied).

1)Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009 Oct;18(8):1115-23. (Slight modifications of the original filters have been performed in order to optimise the search strategy of the current study.

Supplementary Table B1: Identified questionnaires (n=44)

1	AIMS 1	AIMS 2	BASDAI	BASFI	DASH	DGI
2	EQ-5D-3L	FACIT-Fatigue	HAQ-DI	HAQ-SK	HAQ-S	IPBOD
3	mHAQ	MultiP	NRS global*	NRS global (joints)*	NRS global (joints)**	NRS global (skin)*
4	NRS fatigue	Worst itch NRS	NRS pain	PAIP	PASE	PSI
5	PSD	PsAID-9	PsAID-12	$PsAID_{touch}$	PsAQoL	Psodisk
6	SASPA	SF-6D	SF-36	SRPQ	VAS global*	VAS global***
7	VAS global (joints)*	VAS global (skin)*	VAS fatigue	VAS pain	VAS sleep	VITACORA-19
8	WPS	WTP				

AIMS, Arthritis Impact Measurement Scales; BASDAI, Bath Ankylosing Spondylitis Activity Index; BASFI, Bath Ankylosing Spondylitis Functional index; DASH, Disabilities of the Arm, Shoulder and Hand Outcome Measure; DFI, Dougados Functional Index; EQ-5D-3L, EuroQoL 5 Dimensions questionnaire with 3 response levels; FACIT-Fatigue, Functional Assessment of Chronic Illness Therapy-Fatigue Scale; F/C, HAQ-DI, Health Assessment Questionnaire Disability Index, HAQ-S: HAQ Spondyloarthropathy, HAQ-SK: HAQ Skin; IPBOD, Inverse Psoriasis Burden of Disease questionnaire; MultiP, Multidimensional Patient Reported Outcome Questionnaire; NRS, Numeric Rating Scale; NS, Not stated; PAIP, Psoriatic Arthritis Impact Profile; PASE, PsA Screening and Evaluation Questionnaire; PsAID, Psoriatic Arthritis Impact of Disease questionnaire; PsAQoL, PsA Quality of Life instrument; PSD; Psoriasis symptom diary; PSI, Psoriasis Symptom Inventory; PsoDisk questionnaire, no full spelling available; SASPA, Stockerau Activity Score for Psoriatic Arthritis; SF-6D, utility tool derived from SF-36 comprising six multi-level dimensions; SF-36, Medical Outcome Survey Short Form 36-item Health Survey; SRPQ, Social Role Participation Questionnaire; VAS, Visual Analogue Scale; VITACORA-19, Spanish acronym, full name not available; WTP, Willingness to pay questionnaire; WPS, Work Productivity Survey. *1 week recall. ** 1 day recall. *** recall not stated

Supplementary Table B2: Characteristics of the included measurements

PROMs listed by COS domains*	Scales, items, scoring, recall time	Description of PROM (items, subscales)	Develop-ed for	Feasibility, availability and links to more information
MSK DISEASE ACTIVITY				
Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) [31,37,39]	**Scales and items:** 1 scale (6 items). Rating by NRS (0–10) or VAS (0–100 mm). Anchors: "none" and "very severe." Stiffness rated as hours (0 to >2). **Scoring**: The scores for severity and duration of morning stiffness are averaged before calculating the total average score (0-10). No subscale score. Higher BASDAI scores indicate worse disease activity. **Recall time:** 1 week.	The BASDAI includes items of fatigue, pain, swelling, tenderness and stiffness. Generates 1 total disease activity score.	AS	**Completion:** <2 min. **Ease of scoring:** Easy. **Availability**: Free of charge to academic users but not to industry. **More info:** www.asas-group.org. http://oml.eular.org
Stockerau Activity Score for PsA (SASPA)[65]	**Scales and items:** 1 scale (5 items), modified from the RADAI-5 (RA Disease Activity Index). Rating by a 0-10 scale. **Scoring:** Total score is the average of item scores. **Recall time:** 1 day.	Includes items on pain, swelling, global health, morning stiffness, skin disease. Generates 1 total disease activity score.	PsA	**Completion:** <2 min. **Ease of scoring**: Easy **Availability**: Free to use for daily purpose. Copyright: own by Dr.PM Handl and Dr. B Leeb: Burkhard.Leeb@stockerau.lknoe.at
PsA Screening and Evaluation Questionnaire (PASE)[36,70]	**Scales and items**: Two subscales (15 items in total). Rating by a 5-point Likert scale, anchors "strongly disagree" and "strongly agree". **Scoring**: Maximal scale scores: 35 (symptom scale) and 40 (function scale). Total score (max 75). **Recall time:** 1 day	A PsA screening and *evaluation* tool. Symptom scale includes items of pain, fatigue, burning sensation and swelling. Function scale includes work ability, self-care, mobility, physical function and stiffness. 2 sub-scores and 1 total disease activity score are generated.	PsA (and psoriasis)	**Completion:** < 5 min **Ease of scoring:** Easy **Availability:** Copyright: Brigham and Women's Hospital
Multi-P. Patient Reported-Tender	**Scales and items:** One diagram with 76 boxes representing peripheral joints. This PROM is	Boxes corresponding to painful joints are ticked by the patient	RA,PsA, IBD	**Completion:** < 5 min **Ease of scoring:** Easy

joint count (PR-TJC) (72)[44]	part of the Multidimensional "MultiP" described below. **Scoring**: Total tender joint count **Recall**: 1 day	and the total PR-TJC is summed.	arthritis	**Availability:** From the MultiP Questionnaire (described above)
SKIN DISEASE ACTIVITY				
Psoriasis Symptom Inventory (PSI)[67]	**Scale and items:** 1 scale (8 items), rating by a 5-point scale (0-4). **Scoring**: Total score is the sum of the 8 item ratings (0-32). **Recall time:** 1-7 days	Includes items on itch, redness, scaling, burning, stinging, cracking, flaking and pain. Generates 1 total skin disease activity score.	Psoriasis	**Completion:** < 3 min **Ease of scoring:** Easy **Availability**: Development of PSI was sponsored by AMGEN, fee for use not clarified.
Psoriasis Symptom Diary (PSD)[61]	**Scale and items:** 20 NRS scales/items each rated 0-4. **Scoring**: Each scale scored separately **Recall time:** 1 day	Includes items assessing the severity and bother of psoriasis related symptoms and impact.	Psoriasis	**Completion:** < 5 min **Ease of scoring:** Easy **Availability**: for info contact chad.gwaltney@ert.com
Numeric Rating Scale (NRS) of itch[66]	**Scales and items:** 1 scale/1 item, rating by a NRS (0-10). **Scoring:** Higher scores reflect worse itching. **Recall time:** 1 day	A single item generating 1 total itch score (itch related to psoriasis activity)	Psoriasis	**Completion** <1 min **Ease of scoring:** Easy **Availability:** Free of use. Correspondence to: naegelian@lilly.com
PAIN				
Visual Analogue Scale of pain [43,45,50]	**Scales and items:** 1 scale/1 item, rating by a VAS (0-100 mm) **Scoring**: Higher scores reflect worse pain. A score of 0 is "no pain". **Recall time**: 1 week (or not stated)	A single item generating 1 pain score.	PsA/ Generic	**Completion** <1 min **Ease of scoring:** Easy **Availability**: Free of use
Arthritis Measurement Impact Scale versions 1 and 2 (AIMS 1/2) of pain[21,24,25,27, 28]	**Scales and items:** AIMS 1: 1 subscale (4 items) scored on a 6 point VRS (1-6). AIMS2: 1 subscale (5 items) scores on a 5 point VRS (1-5). **Scoring:** Scores within the subscale are summed and a recoding and normalization procedure is performed to gain scores (0-10) (higher scores indicate worse pain) **Recall time:** 1 months	AIMS 1: 4 items on the severity and distribution of pain and stiffness. AIMS 2: 5 items on the severity, frequency, distribution and duration of pain and stiffness and impact on sleep.	RA/OA	**Completion** <1 min **Ease of scoring:** Difficult **Availability:** See description of AIMS below
Numeric Rating	**Scales and items:** 1 NRS (0-10)	1 item: "How much pain have you	PsA/RA/	**Completion** < 1 min

Scale (NRS) of pain[44]	**Scoring:** Higher scores indicate worse pain **Recall time:** 1 week	had because of your arthritis over the past week"	IBD arthritis	**Ease of scoring**: Easy **Availability:** See MultiP below
SF-36 Bodily Pain (BP)[26,28,33,34, 40,50]	**Scales and items:** 1 subscale, 2 items (SF-36 item 7 and 8) scores on a 6 and 5 point VRS. **Scoring**: See description of SF-36 below. **Recall**: 4 weeks	1 item of pain magnitude and 1 item of pain interference on normal activities/work, 1 total score.	Generic	**Completion** < 1 min for this scale **Ease of scoring:** Difficult **Availability:** See SF-36 below
VAS Pain (assessed with HAQ)[22,28]	**Scales and items:** 1 scale/1 item, rating VAS (0-100 mm) (no pain=0 and severe pain =100) **Scoring:** A pain score is calculated by measuring the distance (cm) from 0 to the respondent's mark of pain severity on the line, and multiply with 0.2 to obtain a value from 0-3. **Recall time**: 1 week	1 item, 1 total score.	RA	**Completion** <1 min **Ease of scoring:** Easy-moderate **Availability**: Free of use
PATIENT GLOBAL				
Due to psoriasis				
Patient Global Assessment of skin impact by Numeric Rating Scale[64]	**Scales and items:** 1 scale/1 item, rating by a NRS (0-10). **Scoring:** Higher scores reflect worse global health due to psoriasis **Recall time:** 1 week	1 item: "Considering all the ways psoriasis (skin disease) affected you during the last week, circle the number that best describes how you have been doing"	PsA	**Completion** <1 min **Ease of scoring:** Easy **Availability**: Free of use
Patient Global Assessment of skin impact by Visual analogue Scale[49]	**Scales and items:** 1 scale/1 item, rating by a VAS (0-100 mm) **Scoring**: Higher scores reflect worse global health.due to psoriasis **Recall time:** 1 week.	1 item: "In all the ways your PSORIASIS affects you, how would you rate the way you felt over the past week"	PsA	**Completion** <1 min **Ease of scoring:** Easy **Availability**: Free of use
Due to arthritis				
Patient Global Assessment of joint impact by Numeric Rating Scale[44,64]	**Scales and items:** 1 scale/1 item, rating by a NRS (0-10). **Scoring:** Higher scores reflect worse global health due to PsA joint disease **Recall time:** 1 week or 1 day	1 item: "considering all the ways, your joints affected you during the last week, circle the number that best describes how you have been doing"	PsA	**Completion** <1 min **Ease of scoring:** Easy **Availability**: Free of use
Patient Global Assessment of	**Scales and items:** 1 scale/1 item, rating by a VAS (0-100 mm)	1 item: "In all the ways your ARTHRITIS affects you, how	PsA	**Completion** <1 min **Ease of scoring:** Easy

joint impact by Visual Analogue Scale[49]	**Scoring**: Higher scores reflect worse global health due to PsA joint disease **Recall time:** 1 week.	would you rate the way you felt over the past week."		**Availability**: Free of use
Due to Psoriatic Arthritis				
Patient Global Assessment of PsA impact by Visual Analogue Scale[43,45,49,50,63]	**Scales and items:** 1 scale/1 item, rating by a VAS (0-100 mm). **Scoring**: Higher scores reflect worse global health. **Recall time:** 1 week (most often).	1 item. Example of wording: "In all the ways your PSORIASIS and ARTHRITIS, as a whole, affects you, how would you rate the way you felt over the past week"	PsA/ Generic	**Completion** <1 min **Ease of scoring:** Easy **Availability**: Free of use
Patient Global Assessment of PsA impact by Numeric Rating Scale[53,64]	**Scales and items:** 1 scale/1 item, rating by a NRS (0-10). **Scoring:** Higher scores reflect worse global health. **Recall time:** 1 week	1 item: "Considering all the ways PsA affected you during the last week, circle the number that best describes how you have been doing"	PsA/ Generic	**Completion** <1 min **Ease of scoring:** Easy **Availability**: Free of use
PHYSICAL FUNCTION				
Dougados Functional index (DFI)[34]	**Scales and items**: 1 scale (20 items). 3 point verbal response scale (each item scored 0-2), higher scores reflect worse function. **Scoring:** Total score is the sum of item scores (0–40). **Recall time**: NS. "Usual abilities".	Includes items on physical and daily activities, mobility, and ability to care for one self, turn in bed, breathe deeply and cough. Generates one total score of physical function.	AS/ AxSpA	**Completion**: <3 min. **Ease of scoring**: Easy **Availability:** online (in multiple translations). **More info:** Correspondence to: Prof. M. Dougados: maxime.dougados@aphp.fr
Disability of Arm, Shoulder and Hand Outcome Measure (DASH)[29]	**Scales and items**: 1 scale (30 items). Rating by 5-point scales. Additional 2 optional scales (4 items each). **Scoring:** Formula for calculating total score is available in user's manual. Total score range 0-100, higher scores indicate worse function. **Recall time**: 1 week	Includes items on physical and daily activities, mobility, dexterity, participation in work, social and leisure activities, sleep and sexual problems, pain, weakness and stiffness. Generates 1 total symptom/disability score. 2 optional scales can be applied to assess participation in work	RA, OA, distal radius fracture	**Completion:** <5 min. **Ease of scoring:** Moderate **Availability:** Copyright: www.dash.iwh.on.ca/. Free of charge for non-commercial use; license for commercial use. **More info:** www.dash.iwh.on.ca/

		and sports/arts activities, these generate 2 separate scores.		
Bath Ankylosing Spondylitis Functional Index (BASFI)[34]	**Scales and items**: 1 scale (10 items). Rating by NRS (0-10) or VAS (0–10 cm) Anchors: "easy" and "impossible." **Scoring:** The mean of the 10 item scores provides the overall index score (0-10). **Recall time**: 1 week.	Includes items on physical and daily activities and the ability to care for one self. Generates 1 total score.	AS	**Completion**: <3 min. **Ease of scoring**: Easy **Availability**: Free of charge to academic users but not industry. **More info:** www.asas-group.org. http://oml.eular.org
Health Assessment Questionnaire – Disability Index (HAQ-DI)[22,27,28,33,34,42,43,50,51,60]	**Scales and items:** 1 scale (20 items) of 8 categories of function. 2 subscales: VAS pain and VAS global (0-100mm). Each HAQ-DI item is rated 0-3 (higher scores reflect worse disability). **Scoring:** The highest score within a category is used to calculate the mean score of all categories (total score 0-3). Dependence on physical assistance or equipment raises a category score to 2. VAS scored separately. **Recall time**: 1 week.	Includes items on physical function categorized in 8 areas: Dressing & grooming, arising, eating, walking, hygiene, reaching, gripping, common activities. Generates 1 disability score. Separate additional scales of Patient Global and Pain are often presented with HAQ.	RA	**Completion**: <10 min. **Ease of scoring:** Moderate **Availability:** Copyrighted by Stanford University. There is no charge from Stanford for permission to use HAQ. The HAQ 20-item disability scale is available at http://patienteducation.stanford.edu/research/haq20.html. **More info:** http://oml.eular.org
Health Assessment Questionnaire– Skin (HAQ-SK)[23]	**Scales and items:** 1 scale (23 items), with 3 items added to HAQ-DI to assess skin related disability. 2 subscales: VAS global and VAS pain (0-100mm). **Scoring:** As for the original HAQ-DI. **Recall time:** 1 week	Includes the same items as HAQ (described above) plus 3 items on physical function in relation to psoriasis. Separate additional scales of Patient Global and Pain are often presented with HAQ	RA, skin items for PsA	**Completion:** <10 min. **Ease of scoring:** Moderate **Availability:** NS **More info:** See reference.
Health Assessment Questionnaire – Spondylo-arthropathy (HAQ-S)[22]	**Scales and items:** 1 scale (25 items) with 5 items added to the original HAQ-DI to assess spondylitis related disabilities. 2 subscales: Stiffness and Pain (VAS: 0-100 mm) **Scoring:** As for the original HAQ-DI. **Recall time**: 1 week.	Includes the same items as HAQ (described above) plus 5 items on physical function in relation to spondylitis. Separate additional scales of Stiffness and Pain are presented with HAQ-S.	RA (SpA items for AS)	**Completion:** <10 min. **Ease of scoring:** Moderate **Availability:** NS **More info:** http://oml.eular.org

Modified Health Assessment Questionnaire (mHAQ)[45]	**Scales and items:** This scale is modified from HAQ to include only 8 questions **Scoring:** The score of mHAQ ranges from 0-3. **Recall:** 1 week.	Includes 1 item from each of the 8 areas of physical function presented in the HAQ-DI (described above).		**Completion:** <5 min. **Ease of scoring:** Moderate **Availability:** NS **More info:** http://oml.eular.org, t.pincus@rhul.ac.uk
SF-36 Physical Function Scale (SF-36 PF) [26,28,33,34,40,50]	**Scales and items:** 1 subscale (10 items) from the SF-36 questionnaire (described below). **Scoring:** Item ratings (raw scores) are summed and transformed to obtain a 0-100 scale score. **Recall time:** 1 month.	Includes 10 items assessing different levels of physical and daily activities. Generates 1 physical function score.	Mixed population	**Completion**: < 3 min. **Ease of scoring:** Difficult **Availability and more info:** As for SF-36 (described below)
SF-36 Physical Component Summary (PCS) [40,50]	**Scales and items:** The PCS of SF-36 is derived as an aggregate of the 8 subscale scores. **Scoring:** Z-scores are determined for each of the 8 scale scores and these are multiplied by a factor scoring coefficient and subsequently summed. **Recall time:** 1 month.	Based on the 8 SF-36 subscale scores. Generates one aggregate of physical function.	Mixed population	**Completion, scoring, availability etc.** As for the SF-36 (described below.
MultiP Combined Inflammatory Arthritis – Functional Impairment questionnaire (CIAQ-FI)[44]	**Scales and items:** 1 scale (10 items), VRS (0-3). Part of the MultiP Questionnaire (described below) **Scoring:** Average of the 10 scores with higher score representing worse function. **Recall time:** 1 week	Includes 10 items assessing the difficulty of performing activities of daily living. Generates 1 score of function	RA, PsA IBD-arthritis	**Completion**: < 5 min. **Ease of scoring:** Easy **Availability and more info:** Freely available for clinicians and industry
AIMS1 and AIMS2 Physical Function (AIMS physical) [21,24,25,27,28]	**Scales and items:** The AIMS1 contains 4 subscales of physical/daily function (4-7 items per scale). VRS (2-3 categories). The AIMS2 contains 6 subscales of physical/daily function (4-5 items per scale). 5 point VRS (1-5). **Scoring:** Each function scale is scored separately as for the overall AIMS (described below). **Recall time:** 1 month	Physical function scales of: **AIMS 1 and 2:** Mobility (getting around) Walking/Bending Hand/Finger function Household Selfcare/ADL **AIMS 2:** Arm Function	RA/OA	**Completion:** <2 min. per scale **Ease of scoring:** Difficult **Availability and more info:** Free (see info for the full AIMS below)

		A separate score is generated for each scale.		
AIMS2 Physical component score (AIMS PC)[25,28]	**Scales, items, scoring:** The Physical component score is the average of the 6 AIMS2 function subscale scores (described above). **Recall**: 1 month	Scores from following scales are averaged: Mobility, walking/bending, hand/finger function, arm function, self care, and household tasks.	RA/OA	**Completion:** < 10 min **Ease of scoring:** Difficult **Availability and more info:** Free (see info for the full AIMS below)
HRQoL/Life impact				
Psoriatic Arthritis Quality of Life instrument (PsAQoL) [30,35,41,42,55,71]	**Scales and items:** 1 scale (20 items). Rating: 'true' or 'false' for each item. **Scoring**: Total number of 'true' responses. Higher scores indicate poorer QoL. **Recall time**: 1 day	Includes items on emotional well-being, participation, fatigue, independence and stiffness. Generates 1 total HRQoL score.	PsA	**Completion**: <5 min **Ease of scoring**: Easy **Availability:** Use with permission for a cost: smckenna@galen-research.com
Arthritis Impact Measurement Scale (AIMS1) [21,25,27]	**Scales and items:** 66 items in total. The first 45 items are broken down into 9 subscales (4-7 items per scale). Additional items cover general estimates of health status/health perceptions, overall arthritis impact, medication, comorbidity, and demographics. **Scoring**: A recoding and normalization procedure is needed to express all 9 subscale scores in the range of 0-10. Higher score reflect worse disease impact. Scoring manual available. Summary Scores: A physical, psychological and Pain component score can be calculated by lumping scale scores from similar domains. A manual is available. **Recall time**: 1 month	Assess disease impact by following 9 subscales: 1) Mobility 2) Physical activity 3) Dexterity 4) Household activities 5) Social activities 6) Activities of daily living 7) Pain 8) Depression 9) Anxiety All scales are scored separately but a global AIMS-1 score can be generated as the mean of the 9 scores.	RA/OA	**Completion:** <20 min. **Ease of scoring:** Difficult **Availability:** Free access. Correspondence to: Robert Meenan: rmeenan@bu.edu **More info:** www.proqolid.org/ http://oml.eular.org
Arthritis Impact Measurement Scale-2 (AIMS2) [24,25,28]	**Scales and items:** 78 items with the first 57 items broken down into 12 subscales (4-5 items per scale). Additional items assess satisfaction, health perceptions, arthritis impact, general	Assess the impact of disease by following subscales: 1) Mobility level 2) Physical activity	RA/OA	**Completion**: <20 min. **Ease of scoring:** Difficult **Availability**: Free access, Correspondence to: Robert

	health perception, medication and arthritis/co-morbidity, demographics. **Scoring:** A recoding and normalization procedure is needed to calculate scale scores (0-10). Higher scores reflect worse disease impact. Scoring manual available. **Summary scores**: Factor analysis has suggested a 3 and a 5 component model which group the AIMS measures into general categories. The 3 component model measures Physical Function, Psychological Status and Pain. **Recall time**: 1 month	3) Dexterity 4) Arm function 5) Self-care 6) Household tasks 7) Social activities 8) Social support 9) Arthritis pain 10)Work ability 11)Level of tension 12)Mood All scales are scored separately.		Meenan: rmeenanbu.edu **More info:** www.proqolid.org/ http://oml.eular.org https://eprovide.mapi-trust.org/instruments/arthritis-impact-measurement-scales
PsA Impact of Disease Questionnaire (PsAID-9 and PsAID-12) [58,68,72,73]	**Scales and items:** Two versions: 1 scale (12 items) for routine care and 1 scale (9 items) for trials. Rating by NRS (0-10). **Scoring:** Item scores are multiplied by a weighing score. Sum of the final item scores yield a total score from 0-10, higher scores reflect worse impact. **Recall time**: 1 week	Assesses the impact of disease and includes items on pain, fatigue, skin disease activity, participation (work/leisure), physical function, sleep and emotional well-being. Generates 1 total score.	PsA	**Completion:** <5 min **Ease of scoring:** Moderate **Availability:** English as well as translated versions and scoring instructions freely available. **More info:** http://oml.eular.org
Psoriatic Arthritis Impact Profile (PAIP)[56]	**Scales and items:** 1 scale (23 items) with "4 special parts". Rating by a 4-point scale, higher scores reflect worse disease impact. **Scoring:** Total score is the sum of the scores in the 4 special parts and range from 0-84, higher score indicate worse impact. **Recall time:** NS	Assesses impact of disease and includes items on physical function, emotional well-being, sleep, pain, participation, independence and socio-economic impact of disease. Generates 1 total score. Furthermore, PAIP includes items on demographics, treatment attitude and side effects.	PsA (psoria-sis)	**Completion:** <15 min **Ease of scoring:** Moderate **Availability:** Developed in Italian, non-validated translation of PAIP available in the reference. Correspondence to: Stefano.coaccioli@uniog.it
VITACORA-19[59,69]	**Scales and items:** 1 scale, (19 items), rating by a 5 point Likert scale ("always" to "never"). **Scoring:** Summed score from 0 (worst HRQoL) to	Assesses HRQoL and includes items on physical function, pain, fatigue, participation, emotional-	PsA	**Completion**: < 10 min **Ease of scoring**: Easy **Availability:** Spanish version

	100 (best HRQoL). **Recall time**: 1 week.	wellbeing, disease activity, inflammation, sleep, independence and economy. Generates 1 total score.		available from the author jctorre@telecable.es English version not validated
Psodisk[62]	**Scales and items**: 1 scale (10 items), rating by VAS (0-100 mm), anchors "absolutely no" and "definitely yes". **Scoring:** Scores are joined by a line forming a polygon. A large polygon equals a low quality of life, and decrease of disease burden is visualised by a shrinking of the polygon. **Recall time:** 1 week.	Assesses HRQoL and includes items on physical function, global health, emotional well-being, fatigue, participation, sleep, pain, joint and skin disease activity and economic costs. Generates 1 total "score".	PsA	**Completion:** <5 min **Ease of scoring**: Easy **Availability:** AbbVie sponsored the PsoDisk and made it freely available (as an APP). **More info:** Priv. - Doz. Mag. Dr. Michael Dennis Linder Adjunct Professor, Medical University of Graz, Graz, Austria
Multi-dimensional Questionnaire for PROMs (MultiP)[44]	**Scales and items:** 9 subscales (77 items). Different rating options for the subscales (NRS (0-10)/ 3- point Likert/ joint diagram) **Scoring**: Mean scores for each subscale are calculated. **Recall time**: Current/past week/past month.	Assess disease impact on life by following subscales: 1) Physical function (10 items) 2) Quality of life (10 items) 3) Pain (1 item) 4) Fatigue (1 item) 5) Global health (1 item) 6) Stiffness (duration) (1 item) 7) Joint tenderness (1 diagram) 8) Disease attitude (10 items) 9) Co-morbidities (43 items) Separate scores for each scale are calculated. The PROM also contains general information including co-morbidities and medication.	Arthritis (RA, PsA, IBD-related)	**Completion:** <15 min (mean(SD) completion time was 8.25(0.25) minutes according to reference) **Ease of scoring:** Moderate **Availability:** There are no cost for using it, whether clinician or industry.
MultiP Combined Inflammatory Arthritis – Quality	**Scales and items:** 1 scale, 10 items **Scoring**: 4 point VRS (scores are 0-3), total scale score is the average of items score (0-3) and	10 items concerning ability to get a good night's sleep (1 item), and 9 items concerning ability to cope	Arthritis (RA, PsA, IBD-	**Completion:** <5 min for this scale **Ease of scoring:** Easy **Availability:** There are no cost for

of Life (CIAQ-QoL)[44]	higher scores represent worse quality of life **Recall time**: 1 week	with stressors, social activities, feelings/anxiety, low self-esteem/feeling blue, get going in the morning, usual work, worries about future, continuing usual activities, relationship with partner.	related)	using it, whether clinician or industry.
The 36 item Short Form Health Survey (SF-36) [26,28,33,34,40,50]	**Scales and items:** Different versions of SF-36 have been developed. All versions have 8 subscales (35 items), which each generate separate scale scores. There are 36 items in total one item concerns the change in health. A mental and a physical summary score can be calculated from the subscales. The rating options vary between items (verbal or numeric scales) **Scoring**. The RAND, MOS and SF-36V2 present minor differences in rating and scoring. The most important differences are seen for the Role physical and Role emotional items where the SF-36V2 has 5 response options compared to dichotomous response options in former versions. Raw scores are transformed by a scoring key into values between 0-100 and subscale scores are derived by averaging the values of the items included in the particular scale. Higher scores reflect better health state. A scoring manual is available, including norm-based scoring algorithms. Different normative databases also exist. **Recall time:** 4 weeks.	Assesses HRQoL by 3 overall dimensions: Functional status, Emotional well-being and General health perceptions). Eight separate scales are presented and generate 8 scale scores: 1)Physical function (daily and vigorous activities) 2)Role limitation due to physical health 3) Bodily pain 4) Social functioning 5) General mental health 6) Role limitations due to emotional 7)Vitality (energy/fatigue) 8) General health perception	Mixed popula-tion	**Completion:** <15 min. **Ease of scoring:** Difficult **Availability:** The MOS/RAND 36-Item Short-Form Health Survey (SF-36) is free of charge, while an annual licence fee applies to the SF-36v2. **More info:** http://oml.eular.org http://www.sf-36.com/ https://www.rand.org/health/surveys_tools/mos/36-item-short-form.html

Inverse Psoriasis Burden of Disease questionnaire (IPBOD)[74]	**Scales and items:** 7 items on general information and 16 VAS (0-100 mm) items with anchors "never" and "all the time" referring to how often a given problem/symptom has interfered. **Scoring**: The VAS items are averaged to yield total scale score and/or 5 subscales with 1-4 items in each. **Recall time:** None	Assesses the overall burden of inverse psoriasis by VAS scoring of how much of the time following conditions have been present or affected: Itch, cracking, skin maceration, odor, intimacy, body self-image, shame, physical contact, clothing choices, personal hygiene, school/work, recreational activities, pain, close relationships, depression/anxiety, finances.	PsA PsO	**Completion:** < 10 min **Ease of scoring**: Easy **Availability**: Free to academic users but contact to developers required. Copyright held by Joseph F. Merola, MD MMSc Brigham and Women's Hospital Harvard Medical School. JFMEROLA@BWH.HARVARD.EDU
FATIGUE				
Functional Assessment of Chronic Illness Therapy Fatigue scale (FACIT-Fatigue)[32]	**Scale and items:** 1 scale (13 items) rating by a 5-point Likert scale (0-4). **Scoring**: Items scores are summed to generate a total scale score (0- 52). Higher scores reflect more fatigue. **Recall time:** 1 week	Assesses fatigue by items on tiredness, ability to do/start/finish activities, energy, need for help, participation in social life and frustration. Generates 1 total score.	Cancer Anaemia	**Completion:** <5 min **Ease of scoring**: Easy **Availability:** English version is free to use, a fee is payable for non-English versions in commercial studies. http://www.facit.org/
Numeric Rating Scale (NRS) of fatigue [38,44]	**Scales and items:** 1 scale/1 item, rating by a NRS (0-10). **Scoring:** Higher scores reflect worse fatigue. **Recall time**: NS	Includes 1 item of fatigue generating 1 single score.	RA/PsA Generic	**Completion:** <1 min **Ease of scoring**: Easy **Availability**: Free of use
Visual Analogue Scale (VAS) of fatigue[43]	**Scales and items:** 1 scale/1 item rating by a VAS (0-100 mm). **Scoring**: Higher scores reflect worse fatigue. **Recall time**: 1 week	Includes 1 item of fatigue generating 1 single score.	PsA/ Generic	**Completion** <1 min **Ease of scoring:** Easy **Availability**: Free of use
SF-36 Vitality (SF-36 VT) [26,40,50]	**Scales and items:** 1 scale (4 items) **Scoring**: VRS (1-6) with higher scores reflect more vitality (less fatigue). Item scores are summed and recoded 0-100 (manual available) **Recall**: 1 months	Includes 4 items concerning feeling full of life, having energy, being worn out, and feeling tired.	Mixed populati on	**Completion**: < 3 min. **Ease of scoring:** Difficult **Availability and more info:** As for SF-36 (described above)

PARTICIPATION				
Social Role Participation Questionnaire (SRPQ) (including the 3 scales: SRPQ-IM,SRPQ-ST, SRPQ-SR)[52]	**Scales and items**: 3 subscales (12 items per scale) rating by a 5-point Likert scale. **Scoring:** Mean scores are calculated separately for each of the 3 scales if a person respond to at least 9/12 domains. **Recall period**: Today	Assesses participation by the following 3 scales: 1) Importance of participation 2) Restrictions in role participation 3) Satisfaction in social performance. Separate scores are calculated for each scale.	OA	**Completion**: 10 min **Ease of scoring**: Moderate **Availability**: Free of charge for use in research or clinic, a fee may apply for commercial use/ trial. Copyright owner: M.Gignac and colleges: mgignac@iwh.on.ca
Work Productivity Survey (WPS) [57]	**Scale and items:** 8 subscales (8 items), 12 items in total. The scales/items are rated as "number of days" or by a 0-10 rating scale, anchors "no interference" and "complete interference" (of PsA on productivity) **Scoring:** Each of the scales are scored individually. **Recall time:** 1 month.	Assesses work productivity by following scales: 1) Missed work 2) Reduced work productivity 3) Interference of PsA on work 4) Missed household work 5) Reduced household productivity 6)Missed family activities 7) Need for hiring outside help 8) Interference of PsA on household Each scale generates a separate score. The PROM also includes information on employment status and occupation.	RA	**Completion:** <5 min **Ease of scoring**: Easy **Availability:** The WPS for PsA is originally developed for RA (copyright licence Pharmacia/Pfizer).
AIMS1 and AIMS2 Social Activity (SA) Scale [21,24,25,27,28]	**Scale and items: AIMS1**: 1 subscale (4 items) 6 point VRS (1-6). **AIMS2:** 1 subscale (5 items) 5 point VRS (1-5). **Scoring**: Recoding and normalization to a (0-10) scale score with higher scores representing less social activity. **Recall:** 1 month	For both AIMS versions, the social scales assess the frequency of social activities including having visitors/visit others and being on the telephone.		**Completion:** < 5 min for this scale **Easy of scoring and availability:** See description of AIMS

AIMS2 Work[24]	**Scale and items:** 1 subscale (5 items), 5-6 point VRS. Scoring: Recoding and normalization to a (0-10) scale score with higher scores representing less ability to perform work. **Recall:** 1 month	Includes 5 items on the type of work and the ability to work in a normal way, have a full day of work and doing the work as carefully as usual.		**Completion:** < 5 min for this scale **Easy of scoring and availability:** See description of AIMS 2.
AIMS2 Social component (SC) score[25,28]	**Scales, items, scoring:** The Social component score is the average of the 2 social AIMS 2 subscale scores. **Recall**: 1 month	Scores from the social activity and the social support scales are averaged:	RA/OA	**Completion:** < 15 min **Ease of scoring:** Difficult **Availability and more info:** Free (see info for the full AIMS below)
SF-36 Role Emotional (SF-36 RE)[26,40,50]	**Scales and items:** 1 subscale (3 items) **Scoring**: Dichotomous response (1 Yes/2 No) (elaborated to 5 response categories in the SF-36v2). Recoding of these into a 0-100 scale (manual available). **Recall**: 1 months	Includes 3 items concerning the impact of emotional problems on time spent on work/other activities, accomplishing things, doing things carefully.	Mixed population	**Completion**: < 3 min. **Ease of scoring:** Difficult **Availability and more info:** As for SF-36 (described above)
SF-36 Role Physical (SF-36 RP)l[26,40,50]	**Scales and items:** 1 subscale (4 items) **Scoring:** Dichotomous response (1 Yes/2 No) (elaborated to 5 response categories in the SF-36v2). Recoding of these into a 0-100 scale (manual available). **Recall**: 1 months	Includes 4 items concerning the impact of physical problems on time spent on work/other activities, accomplishing things, limitations in work/other activities, difficulties performing work/other activities.	Mixed population	**Completion**: < 3 min. **Ease of scoring:** Difficult **Availability and more info:** As for SF-36 (described above)
SF-36 Social Functioning (SF-36 SF)[26,28,40,50]	**Scales and items:** 1 subscale (2 items) **Scoring:** VRS (1-5) higher values representing more interference of normal social activities (less participation)	Includes 2 items concerning the extent and amount of time that physical and emotional problems interfered with normal social activities.	Mixed population	**Completion**: < 3 min. **Ease of scoring:** Difficult **Availability and more info:** As for SF-36 (described above)
EMOTIONAL WELL-BEING				

SF-36 Mental Health (SF-36 MH)[26,28,40,50]	**Scale and items:** 1 subscale (5 items), VRS (1-6) **Scoring:** Recoding of items scores into a 0-100 scale (manual available). **Recall:** 1 month	5 items addressing nervousness, and the presence of feeling "down in the dumps", peaceful/calm, downhearted, happy.		**Completion**: < 3 min. **Ease of scoring:** Difficult **Availability and more info:** As for SF-36 (described above)
SF-36 Mental Component Summary (SF-36 MCS)[40,50]	**Scales and items:** The MCS of SF-36 is derived as an aggregate of the 8 subscale scores. **Scoring:** Z-scores are determined for each of the 8 scale scores and these are multiplied by a factor scoring coefficient and subsequently summed. **Recall time:** 1 month.	Based on the 8 SF-36 subscale scores. Generates one aggregate of emotional well-being.	Mixed population	**Completion, scoring, availability etc.** As for the SF-36 (described below.
AIMS1 and AIMS2 Anxiety/Tension Scales[21,24,25,27,28]	**Scale and items:** AIMS 1: 1 subscale (6 items), 6-point VRS (1-6). AIMS 2: 1 subscale (5 items), 5 point VRS. **Scoring**: Recoding and standardization to a (0-10) score, higher scores indicate worse anxiety. **Recall:** 1 month	Items of both AIMS1 and AIMS2 concern the frequency of tension/anxiety symptoms (feeling tense, bothered by nervousness, difficulty relaxing, feeling calm)		**Completion:** < 5 min for this scale **Easy of scoring and availability:** See description of AIMS
AIMS1 and AIMS2 Depression/ Mood Scales[21,24,25,27,28]	**Scale and items:** AIMS 1: 1 subscale (6 items), 6-point VRS (1-6). AIMS 2: 1 subscale (5 items), 5 point VRS. **Scoring**: Recoding and standardization to a (0-10) score, higher scores indicate worse mood. **Recall:** 1 month	Items of both AIMS1 and AIMS2 depression/mood scales concern the frequency of depressive sym-ptoms (enjoying things, feeling low in spirits, feeling nothing turned out right, down in dumps)		**Completion:** < 5 min for this scale **Easy of scoring and availability:** See description of AIMS
AIMS2 psychological component (Psyc.C) score [25,28]	**Scale, items and scoring:** This component score is the average of the anxiety/tension and the Mood/Depression subscale scores. **Recall**: 1 month	See description of the Tension and the Mood AIMS scales for more information on content.		**Completion:** < 10 min **Easy of scoring and availability:** See description of AIMS

MultiP Modified Rheumatology Attitude Index (mRAI)[44]	**Scales and items:** 1 scale, 10 items **Scoring**: 10 point NRS, higher scores represent worse emotional well-being. **Recall**: 1 week	10 items concerning the presence of worries (related to the disease.)	Arthritis (RA, PsA, IBD-related)	**Completion:** <5 min for this scale **Ease of scoring:** Easy **Availability:** There are no cost for using it, whether clinician or industry.
ECONOMIC COST				
EuroQol-5D-3L (EEQ-5D)[42,45,47,48,54,75] EQ-5D-revised[48]	**Scales and items**: 1 descriptive scale (5 items) and 1 subscale (VAS). Rating of descriptive scale: 3 level (EQ-5D-3L) or 5 level (EQ-5D-5L) Likert scale. VAS (0-100 mm), anchors "best" and "worst" health state. **Scoring:** Scores can be converted into a summary (EQ-5D 'Index') that uses a utility-weighted scoring system. EQ-5D VAS scores can be converted into Quality-Adjusted-Life-Year. **Recall time**: Today A revised scoring system for EQ-5D(UK) time trade off, further described by the authors[48]	Assesses HRQoL and includes items on physical function, independence, participation, pain and emotional well-being. This scale generates 1 total score. A subscale of Patient Global health is scored separately.	General pop-ulation	**Completion:** < 10 min **Ease of scoring**: Difficult **Availability**: User fees determined by the EuroQol Executive Office userinformationservice@euroqol.org **More info:** http://www.euroqol.org/eq-5d-products.html (User's guide etc.)
SF-6D[45,47,54]	**Scales and items:** 1 scale/ index (11 items) measuring 6 of the 8 SF-36v2 domains. **Scoring**: The SF-6D index is scored from 0.0 (worst health state) to 1.0 (best health state). **Recall time:** 4 weeks	Includes items SF-36v2 scales: 1)Physical functioning *(3a,3b,3j)* 2)Role participation (combined RP and RE, 4c*, 5b)* 3)Social functioning *(10)* 4) Bodily pain *(7,8)* 5)Mental health *(9b, 9f)* 6)Vitality *(9e)* A single score is generated.	Mixed popu-lation	**Completion:** <3 minutes **Ease of scoring:** Difficult **Availability:** For commercial applications there is a per study license (https://www.optum.com)
Willingness to pay questionnaire (WTP)[46]	**Scales and items**: 8 subscales (27 items) and 2 VAS (0-100 mm) subscales. Items of the 8 subscales rated by "amount of money willing to pay for a health problem to resolve" **Scoring:** Amount of money willing to pay for resolution of each item provides information on	Assesses health related quality of life by following scales: 1) Intimacy 2) Physical comfort (including aspects of pain and skin symptoms)	PsA (other WTP PROMs pre-viously	**Completion**: <30 min **Ease of scoring**: Moderate **Availability**: Reprint request: AA Qureshi, Harvard Medical School, Department of Dermatology, Brigham and Women's Hospital,

	the impact of PsA for each aspect of HRQoL. **Recall time**: None	3) Self-care (physical tasks) 4) Work/Family (participation) 5) Concentration 6) Emotional health 7) Social comfort (participation) 8) Sleep Besides these scales, the PROM includes a subscale of Patient Global Health and general information on demographics, economy and disease characteristics.	used in e.g., psoriasis)	US. aqureshi@bics.bwh.harvard.edu WTP and instructions are depicted in the reference.
SLEEP				
Visual Analogue Scale (VAS) of sleep[43]	**Scales and items:** 1 scale/1 item rating by a VAS (0-100 mm) **Scoring:** Higher scores reflect worse sleep problems. **Recall time**: 1 week	**Main construct:** Sleep **Scale domain:** Sleep	PsA/ Generic	**Completion** <1 min **Ease of scoring:** Easy **Availability**: Free of use
STIFFNESS				
NRS stiffness[44]	**Scales and items:** 1 item (From the multidimensional MultPROM, described above) **Scoring:** Minutes or hours of morning stiffness **Recall:** 1 week	1 item addressing the presence and duration of morning stiffness	Arthritis (RA, PsA, IBD-related)	**Completion:** <1 min for this scale **Ease of scoring:** Easy **Availability:** There are no cost for using it, whether clinician or industry.
HAQ-S VAS stifness[22]	**Scales and items:** 1 scale/1 item rating by a VAS (0-100 mm). **Scoring**: Higher scores reflect worse stiffness. **Recall time**: NS	Includes 1 item of stiffness.	PsA/ Generic	**Completion** <1 min **Ease of scoring:** Easy **Availability**: Free of use
NON-COS DOMAINS				

SF-36 General Health(GH) [26,40,50]	**Scales and items:** 1 subscale (5 items) **Scoring:** 5 point VRS. Scores are transformed to a (0-100) scale (manual available)- **Recall**: Today/"generally"	5 items concerning perception of general health status	Mixed population	**Completion:** <1 min for this scale **Ease of scoring:** Difficult **Availability:** As or SF-36 (described above)
AIMS-2 Social Support[24]	**Scales and items:** 1 subscale (4 items). **Scoring:** Each item is scored on a VRS (1-5). Total scale score is the average of items scores, converted to a 0-10 score, higher scores represent less social support. **Recall: 1 month**	4 items addressing if patients are satisfied with the (frequency of) support, assistance and understanding provided by their friends and family	RA/OA	**Completion:** < 10 min **Easy of scoring and availability:** See description of AIMS

*COS domain: Domains listed and phrased according to the revised PsA "Core Outcome Set". Original target population refers to the population in which the PROM was developed. AS, ankylosing spondylitis; AxSpA, Axial Spondyloarthritis; IBD, Inflammatory bowel disease; Min, Minutes; mm, millimeter; MSK, Muscular skeletal; NS, Not Stated; OA, osteoarthritis; PsA, Psoriatic Arthritis; RA, Rheumatoid arthritis VRS; verbal response scale.

Supplementary Table C: Methodological quality (excellent, good, fair, poor) of each study per measurement property per PROM and scoring of the measurement property results (+/-/?)

Identified PROMs listed according to Domain category	Reliability BOX (A-C)			Validity BOX (D-H)					Responsiveness BOX (I)	Info on score interpretation (values are provided in Suppl. Table D)
	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypotheses testing	Cross-cult. Validity	Criterion validity	Responsiveness, Sens. To change	
	A	B	C	D	E	F	G	H	I	
MSK DISEASE ACTIVITY										
BASDAI *Eng*[31]						Fair/-				F/C
BASDAI *Span*[37]	Poor/?					Fair/-			Poor/?	F/C
BASDAI *Eng*[39]						Fair/+				
SASPA *Germ*[65]	Fair/+				Fair/+	Poor/?			Poor/?	
PASE-total *Eng*[36]		Poor/?				Poor/?			Poor/?	
PASE-symptom *Eng*[36]		Poor/?				Poor/?			Poor/?	
PASE-function *Eng*[36]		Poor/?				Poor/?			Poor/?	
PASE-total *Ital*[70]						Fair/+	*a*		Fair/+	
PASE-symptom *Ital*[70]						Fair/+	*a*		Fair/+	
PASE-function *Ital*[70]						Fair/+	*a*		Fair/+	
PR-TJC *Eng*[44]				Poor/?		Poor/?				
SKIN DISEASE ACTIVITY										
PSI *Eng*[67]	Good/+	Fair/+			Good/+	Fair/+			Fair/+	F/C
PSD[61]				Excellent/+						
NRS ITCH *Eng*[66]				Fair/+						

PAIN						
VAS pain (1 week recall) *Eng*[43]						MID
VAS pain (recall unknown) *Norw*[45]						MCII, PASS
VAS pain (1 week recall) *Chin*[50]					Poor/?	MID
VAS Pain (HAQ, 1 week recall), *Eng*[28]					Poor/?	
VAS Pain (HAQ, 1 week recall), *Eng*[22]				Fair/+		
NRS pain (1 week recall) *Eng*[44]		Poor/?	Poor/?	Poor/?		
SF-36 BP *Eng*[26]	Poor/?			Fair/+		
SF-36 BP *Eng*[28]	Poor/?				Poor/?	
SF-36 BP *Chin*[40]	Poor/?			Good/+*b*		F/C
SF-36 BP *Chin*[50]					Poor/?	MID
AIMS1 Pain *Eng*[21]				Fair/+		
AIMS1 Pain *Eng*[25]					Poor/?	
AIMS1 Pain *Ital*[27]				Fair/+		
AIMS2 Pain, *Eng*[24]				Fair/+		
AIMS2 Pain *Eng*[28]	Poor/?				Poor/?	
AIMS2 Pain *Eng*[25]					Poor/?	

Table C cont.	Reliability			Validity					Responsiveness	Info on score interpretation
Identified PROMs listed according to Domain category	**COSMIN BOX (A-C)**			**COSMIN BOX (D-H)**					**COSMIN BOX: I**	
	Internal consistency	**Reliability**	**Measure-ment error**	**Content validity**	**Structural validity**	**Hypothe-ses testing**	**Cross-cult. Validity**	**Criterion validity**	**Responsiveness, Sens. To change**	**(values are provi-ded in suppl. Table D)**
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
PATIENT GLOBAL										
Patient global (Psoriasis)										
NRS skin (1 week recall) *Eng*[64]						Fair/+				F/C
VAS skin (1 week recall) *Eng*[49]		Good/+				Poor/?				
Patient Global (Arthritis)										
NRS joints (1 week recall) *Eng*[64]						Fair/+				F/C
NRS joints (1 day recall) *Eng*[44]				Poor/?		Poor/?				
VAS joints (1 week recall) *Eng*[49]		Good/+				Poor/?				
Patient Global (PsA)										
PGA by NRS (1 week recall) *Eng*[64]						Fair/+				F/C
PGA by NRS (1 week recall) Chin[53]						Fair/+				
PGA by VAS (1 week recall) *Eng*[43]										MID
PGA by VAS (1 week recall) *Eng*[49]		Good/+				Poor/?				
PGA by VAS (1 week recall) *Ital*[63]						Fair/+				
PGA by VAS (recall unknown) *Norw*[45]										PASS, MCII
PGA by VAS (1 week recall) *Chin*[50]									Poor/?	MID

Table C cont.

PHYSICAL FUNCTION					
DFI *Chin*[34]	Good/-	Good/-	Poor/?		F/C
DASH *Eng*[29]			Good/-		
BASFI *Chin*[34]	Good/+	Good/+	Poor/?		F/C
HAQ-DI *Eng*[22]			Fair/-		
HAQ-DI *Eng*[28]	Poor/?			Poor/?	
HAQ-DI *Eng*[33]	Good/+	Good/+			F/C
HAQ-DI *Eng*[43]					MID
HAQ-DI *Eng*[51]					MID
HAQ-DI *Ital*[27]			Fair/-		
HAQ-DI *Chin*[34]	Good/+	Good/+	Poor/?		F/C
HAQ-DI *Hung*[42]			Fair/+		
HAQ-DI *Chin*[50]				Poor/?	MID
HAQ-DI *Thai*[60]	Poor/?		Fair/+		F/C
HAQ-S *Eng*[22]			Fair/-		
HAQ-SK *Eng*[23]			Poor/?		
mHAQ *Norw*[45]					MCII, PASS
SF-36 PF *Eng*[33]	Good/+	Good/+			F/C
SF-36 PF *Chin*[34]	Good/+	Good/+	Poor/?		F/C
SF-36 PF *Eng*[26]	Poor/?		Fair/+		
SF-36 PF *Eng*[28]	Poor/?			Poor/?	

Table C cont.

SF-36 PF *Chin*[40]	Poor/?				Good/+*b*		F/C
SF-36 PF *Chin*[50]						Poor/?	MID
SF-36 PCS *Chin*[40]				Good/+	Poor/?		
SF-36 PCS *Chin*[50]						Poor/?	MID
CIAQ-FI[44]		Poor/?	Poor/?		Poor/?		
AIMS1 Mobility *Eng*[21]					Fair/-		
AIMS1 Physical *Eng*[21]					Fair/+		
AIMS1 Dexterity. *Eng*[21]					Fair/+		
AIMS1 House *Eng*[21]					Fair/+		
AIMS1 ADL *Eng* [21]					Fair/-		
AIMS1 PC *Eng*[25]						Poor/?	
AIMS1 Physical *Ital*[27]					Fair/-		
AIMS2 PC *Eng*[28]						Poor/?	
AIMS2 Mobility *Eng*[24]					Fair/+		
AIMS2 Physical *Eng*[24]					Fair/+		
AIMS2 Dexterity. *Eng*[24]					Fair/+		
AIMS2 Selfcare *Eng*[24]					Fair/-		
AIMS2 House *Eng*[24]					Fair/-		
AIMS2 Arm F. *Eng*[24]					Fair/+		
AIMS2 PC *Eng*[25]						Poor/?	

Table C cont.	Reliability			Validity					Responsiveness	Info on score interpretation
Identified PROMs listed according to Domain category	**COSMIN BOX (A-C)**			**COSMIN BOX (D-H)**					**COSMIN BOX: I**	
	Internal consistency	**Reliability**	**Measure-ment error**	**Content validity**	**Structural validity**	**Hypothe-ses testing**	**Cross-cult. Validity**	**Criterion validity**	**Responsiveness, Sens. To change**	**(values are provi-ded in Suppl. Table D)**
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
HRQol/Life Impact										
PsAQoL *Eng*[30]	Good/+	Fair/?		Excellent/+	Good/+	Fair/+				
PsAQoL *Eng*[35]						Poor/?			Poor/?	
PsAQol *Eng/Chin*[71]	Poor/?	Fair/+		Poor/?		Fair/+	*a*			
PsAQoL Swe[41]	Good/+	Poor/?		Good/+		Fair/+	*a*			F/C
PsAQoL *Hung*[42]						Fair/+				
PsAQoL *Dutch*[55]	Poor/?	Good/+	Good/?	Good/-		Fair/+	*a*			
AIMS1 Global *Eng*[27]						Poor/?				
PsAID-9 *Eng*[58]	*c*	Good/+				Fair/+	*a*		Poor/?	PASS
PsAID-12 *Eng*[58]	*c*	Good/+				Fair/+	*a*		Poor/?	PASS
PsAID *Eng*[68]				Excellent/+						
PsAID-12 *Ital*[73]	*c*				*c*	Fair/+				Cut off values
PsAID-12touch *Ital*[72]	*c*					Fair/+		Fair/+		MDA Cut-off
PAIP *Ital*[56]						Poor/?				
VITACORA-19 *Span*[59]	Fair/+	Good/+		Good/+	Fair/+	Fair/+			Poor/?	F/C,MCID
VITACORA-19 *Turk*[69]	Poor/?	Fair/+			Poor/?	Fair/+	*a*			

PsoDisk *Ital*[62]						Poor/?	
CIAQ-Qol *Eng*[44]		Poor/?		Poor/?	Poor/?		
IPBOD[74]	Poor/?			Poor/?	Poor/?		
FATIGUE							
FACIT-Fatigue *Eng*[32]	Poor/?	Fair/+			Fair/+		
NRS fatigue (recall NS) *Eng*[38]					Poor/?	Poor/?	
NRS fatigue (recall NS) *Eng*[44]		Poor/?		Poor/?	Poor/?		
VAS fatigue (1 week recall) *Eng*[43]							MID
SF-36 VT *Eng*[26]	Poor/?				Fair/-		
SF-36 VT *Chin*[40]	Poor/?				Good/+*b*		F/C
SF-36 VT *Chin*[50]						Poor/?	MID
PARTICIPATION							
SRPQ-IM *Eng*[52]	Poor/?	Fair/+	Fair/?	Fair/+	Fair/+		MDC
SRPQ-ST *Eng*[52]	Poor/?	Fair/+	Fair/?	Fair/+	Fair/+		MDC
SRPQ-SR *Eng*[52]	Poor/?	Fair/+	Fair/?	Fair/+	Fair/+		MDC
WPS *Eng*[57]					Fair/+	Fair/+	F/C
AIMS1 SA *Eng*[21]					Fair/?*d*		
AIMS2 SA *Eng*[24]					Fair/?*d*		
AIMS2 Work *Eng*[24]					Fair/?*d*		
AIMS2 SC. *Eng*[28]						Poor/?	
SF-36 RE *Eng*[26]	Poor/?				Fair/?		
SF-36 RE *Chin*[40]	Poor/?				Good/+*b*		F/C

SF-36 RE *Chin*[50]						Poor/?	MID
SF-36 RP *Eng*[26]					Fair/-		
SF-36 RP *Chin*[40]	Poor/?				Good/+*b*		F/C
SF-36 RP *Chin*[50]						Poor/?	MID
SF-36 SF *Eng*[26]					Fair/?		
SF-36 SF *Eng*[28]	Poor/?					Poor/?	
SF-36 SF *Chin*[40]	Poor/?				Good/+*b*		F/C
SF-36 SF *Chin*[50]						Poor/?	MID
EMOTIONAL WELL-BEING							
SF-36 MH *Eng*[26]	Poor/?				Fair/?		
SF-36 MH *Eng*[28]	Poor/?					Poor/?	
SF-36 MH *Chin*[40]	Poor/?				Good/+*b*		
SF-36 MH *Chin*[50]						Poor/?	MID
SF-36 MCS *Chin*[40]				Good/+	Poor/?		
SF-36 MCS *Chin*[50]						Poor/?	MID
AIMS1/2 Psyc.C. *Eng*[25]						Poor/?	
AIMS1 Anxiety *Eng*[21]					Fair/?*d*		
AIMS1 Depression *Eng*[21]					Fair/?*d*		
AIMS2 Mood *Eng*[24]					Fair/?*d*		
AIMS2 Tension *Eng*[24]					Fair/?*d*		
AIMS2 Psyc. C. *Eng*[28]						Poor/?	
mRAI (MultiP) *Eng*[44]		Poor/?	Poor/?		Poor/?		

ECONOMIC COST					
EQ-5D-3L *Norw*[45]					PASS, MCII
EQ-5D-3L *Eng*[47]			Poor/?	Poor/?	Score distribution
EQ-5D-3Lrev *Eng*[48]			Poor/?	Poor/?	Score distribution
SF-6D *Eng*[47].			Poor/?	Poor/?	Score distribution
EQ-5D-3L *Swe*[75]					PASS
EQ-5D-3L *Hung*[42]			Fair/+		
EQ-5D-3L *Eng/Chin*[54]			Fair/+		F/C
SF-6D *Eng/Chin*[54]			Fair/+		F/C
SF-6D *Norw*[45]					MCII, PASS
WTP *Eng*[46]		Poor/?	Fair/+		
SLEEP					
VAS sleep (1 week recall) *Eng*[43]					MID
STIFFNESS					
NRS stiffness (1 day recall) *Eng*[44]			Poor/?		
VAS stiffness (assessed with HAQ, 1 week recall) *Eng*[22]			Poor/?		
NON-COS Domains					
SF-36 GH *Eng*[26]	Poor/?		Fair/-		
SF-36 GH *Chin*[40]	Poor/?		Good/-*b*		
SF-36 GH *Chin*[50]				Poor/?	MID
AIMS2 Social Support *Eng*[24]			Fair/?*d*		

Empty cells reflect that the measurement property was not evaluated by any study for the given instrument**.** **T**able 2 explains the grading of evidence (+/-/?).
[a]Only translation, no cross-cultural validation. According to COSMIN, only studies that address measurement invariance (e.g. multiple group factor analyses or

DIF) between countries (or other groups) are considered real cross-cultural validity studies.[b]Construct validity – hypotheses testing was assessed regarding the internal relationships (scale assumptions) and not relationship to external measures [c]Questionnaire based on formative model why internal consistency and structural validity are not rated. [d] Only relations to measures of other constructs presented. AIMS, Arthritis Impact Measurement Scales (ADL, Activity of daily living; Arm F., Arm Function; House, Household; PC, Physical component score; Psyc.C., Psychological component score; SA, Social Activity, SC, Social component score); BASDAI, Bath Ankylosing Spondylitis Activity Index; BASFI, Bath Ankylosing Spondylitis Functional index; Chin, Chinese; CIAQ-FI, Combined Inflammatory Arthritis – Functional Impairment questionnaire; CIAQ-QoL, Combined Inflammatory Arthritis – quality of life questionnaire; COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; DASH, Disabilities of the Arm, Shoulder and Hand questionnaire; DFI, Dougados Functional Index; Eng, English; EQ-5D-3L, EuroQoL 5 Dimensions questionnaire with 3 response levels; FACIT-Fatigue, Functional Assessment of Chronic Illness Therapy-Fatigue Scale; F/C, Floor/Ceiling effect; Germ, German; HAQ, Health Assessment Questionnaire (HAQ-S: Spondyloarthropathy, HAQ-SK: Skin, HAQ-DI: Disability Index); Hung, Hungarian; IPBOD, Inverse Psoriasis Burden of Disease questionnaire Ital, Italian; MCID, Minimal Clinically Important Difference; MDC, minimal detectable change; MCII, Minimal clinical important improvement; MIC, Minimal important change; MID, Minimal important Difference; mRAI, Modified Rheumatology Attitude Index; MultiP, Multidimensional Patient Reported Outcome Questionnaire; Norw, Norwegian; NRS, Numeric Rating Scale; NS, Not stated; PAIP, Psoriatic Arthritis Impact Profile; PASE, PsA Screening and Evaluation Questionnaire; PASS, Patient acceptable symptom state; PGA, Patient Global Assessment; PR-TJC, Patient-reported-tender-joint-count; PsAID, Psoriatic Arthritis Impact of Disease questionnaire; PsAQoL, PsA Quality of Life instrument; PSI, Psoriasis Symptom Inventory; PsoDisk questionnaire, no full spelling available; SASPA, Stockerau Activity Score for Psoriatic Arthritis; SF-6D, utility tool derived from SF-36 comprising six multi-level dimensions; SF-36, Medical Outcome Survey Short Form 36-item Health Survey (SF-36 scales: BP, Bodily Pain; GH, General Health; MCS, Mental Component Summary; MH, Mental Health; PCS, Physical Component Summary, PF, physical function; RE, Role Emotional; RP, Role Physical; SF, Social Functioning; VT, Vitality); Span, Spanish; SRPQ, Social Role Participation Questionnaire; Swe, Swedish; Turk, Turkish; VAS, Visual Analogue Scale; VITACORA-19, Spanish acronym, full name not available; WTP, Willingness to Pay questionnaire; WPS, Work Productivity Survey.

Supplementary Table D: Methodological quality of each study per measurement property (excellent/good/fair/poor) of each instrument assessed, and scoring of the measurement property results (+/-/?)

Identified PROMs listed according to Domain category	Reliability COSMIN BOX (A-C)			Validity COSMIN BOX (D-H)					Responsiveness COSMIN BOX I	Relevant info on score interpretation
	Internal consistency (A)	Reliability (B)	Measurement error (C)	Content validity (D)	Structural validity (E)	Hypotheses testing (F)	Cross-cult. validity (G)	Criterion validity (H)	(I)	
MSK DISEASE ACTIVITY	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
BASDAI *Eng*[31]						Fair/-				F/C
Results: Box F: Vague hypotheses and sparse information about measurement properties of comparators. Group with axial PsA was small (n = 37). BASDAI showed greatest correlation with other PROMs (e.g., PGA r = 0.73) and not with measures of disease activity or damage. No difference in correlation between BASDAI score and patient's perception of arthritis activity in axial vs. peripheral PsA. **Interpretability:** N = 133. Missing data: 16.3%. Median(IQR) score: 2.95(1.50–4.84). Median score peripheral vs axial PsA: 3.07 vs 4.08. Floor effect: 1.3%, Ceiling effect: 0%										
BASDAI *Spanish*[37]	Poor/?					Fair/-			Poor/?	F/C
Results: Box A: Unidimensionality not checked. Cronbach-α = 0.647 (AxPsA) and 0.783 (Peripheral PsA). Box F: Vague hypotheses and sparse information about measurement properties of comparators. The correlation to other measures of disease activity (PGA, PhGA, spinal pain, BASFI, HAQ, SF-36 PF.) was similar in axPsA and pePsA (unexpected). Box I: Not clear if a proportion of patients had changed over time, no correlation between change scores. The change in BASDAI score did not show correlation to disease state at follow up for peripheral or axial PsA. **Interpretability:** N = 100 (Axial and peripheral PsA). Missing data: 0%. Mean(SD) median score: Peripheral PsA: 1.7(1.8),1.2, Axial PsA: 2.7(1.9),2.6. Time frame (responsiveness), mean(SD): 12.1(2.1) months. Floor effect: 33% for peripheral PsA; 14.3% for axial PsA. Ceiling effect: 0%										
BASDAI *Eng*[39]						Fair/+				
Results: Box F: Vague hypotheses and sparse information about properties of comparators however the PGA, PhGA and need for treatment change seem to be fair indicators of disease activity (face validity). In AxPsA the BASDAI correlated highly with PGA (r = 0.81) and moderately with PhGA (r = 0.53) and BASDAI predicted high disease activity measured by: 1) Physician rating: BASDAI OR =1.53, AUC (95%CI): 0.78 (0.67-0.88), 2) Patient rating: BASDAI OR = 2.54, AUC(95%CI): 0.92(0.88-0.95), and 3)Change in treatment: BASDAI R^2 = 1.31, AUC(95%CI): 0.69(0.63-0.76). **Interpretability:** N = 201 (axial PsA). Missing data: max 5.4% (excl.). Mean(SD) score: 3.5(2.4). Floor/ceiling effect: NS.										
SASPA *German*[65]	Fair/+				Fair/+	Poor/?			Poor/?	
Results: Box A, E: Unidimensionality checked. Cronbach-α = 0.875. FA: Sample size <100. Eigenvalue 3.628. Explained variance (3.628/6) = 60%. High factor loadings. Box F,I: No hypotheses and no information about measurement properties of comparators. Statistically significant difference in median SASPA score was seen between different levels of PatSat (defined as "Patient's satisfaction with disease state") but no exact results were provided (only box plots). Only 19 patients in the sensitivity to change analysis, no a priori hypothesis about expected magnitude of effect, but SMD found to be 2.1.I**nterpretability:** N = 152. Missing data: NS. Mean(range) score: 2.66(0-9.2). Mean(range) score for patients undergoing treatment (n										

= 19): Baseline 4.51(1.6-7.2) and after therapy 1.87(0.2-4.4). Time frame (responsiveness), mean: 4.1 months. Floor/Ceiling effect: NS.										
PASE, 3 scales (symptom, function total scale) *Eng*[36]		Poor/?				Poor/?				Poor/?
Results: Box B: Small sample (n = 23) (not clear how many patients had PsA vs Pso) and time interval not clear (> 2 weeks). ICC = 0.9 for the entire PASE Score, not reported for the separate scales. Box F,I: Only assessed by know-group validity approach and without a priori hypotheses. Both the functional, the symptom and the total median PASE scale scores were higher in PsA vs non-PsA patients (p<0.05). Box I: Small sample size (n = 24) in the analysis of responsiveness. PASE scores were significantly different in PsA vs. Psoriasis patients and decreased more in PsA vs Psoriasis patients after treatment. **Interpretability**: N = 37 with PsA (190 in total). Missing data: 2% (excl.). Median(IQR) score (PsA): Functional score: 26(22-30), symptom score: 24(23-27), total score: 51(44-57). Time frame: Test-retest: >2 weeks, responsiveness: 19 weeks (median). Floor/ceiling effect: NS.										
PASE, 3 scales (symptom, function total scale) *Ital[70]*						Fair/+	Poor/?			Fair/+
Results: Internal consistency and reliability only reported for the total population where PsA were less than 50% and were therefore not included in this review (Cronbach-α = 0.90-0.95,Test-retest, ICC = 0.91-0.93) for scales. Box F: Known-group validity and a priori hypothesis confirmed as patients with PsA had significantly higher PASE scores compared to those with psoriasis only however no measures of distribution were presented. For the overall population (PsA <50%) convergent validity was demonstrated with correlation to scores of VAS pain 0.51-0.53. Box G: Only translation no cross-cultural validation. No description of pre-testing (cognitive interview) after translation was reported. Box I: Known group approach showing significant differences in the improvement of PASE scores according to rating of clinical improvement (for the overall population), and patients with PsA diagnosis improved more in PASE scores compared to those without PsA. Hypotheses about excepted differences were vaguely stated. **Interpretability**: N = 298 (PsA n= 28-56). Missing data: NS. Mean(SD) PASE scores: Functional score: 21.3, symptom score: 19.6, total score: 40.9. SD not stated. Time frame responsiveness: 3 months. Floor/ceiling effect: NS.										
PR-TJC *Eng*[44]					Poor/?	Poor/?				
Results: Box D: Not enough information available to rate the study or results regarding content validity. Box F: No information on measurement properties of comparators (Physician assessed 28 TJC), sparse hypotheses and description of the joint diagram. However a strong correlation between the Patient Reported and Physician assessed TJC (r=0.799) was shown. **Interpretability:** Total N = 462 (PsA 123, 26.6%). Number of PsA patients not reported for all analyses. Missing data: NS. Floor/Ceiling effect: NS.										
SKIN DISEASE ACTIVITY	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
PSI *Eng*[67]	Good/+	Fair/+			Good/+	Fair/+			Fair/+	F/C
Results: Box A,E: Unidimensionality checked. Cronbach-α = 0.95 (baseline) and 0.97 (week 12). PCA: Comparative Fit Index>0.90, weighted root mean square residual (WRMR) <1. Rasch: The PSI items exhibited well-ordered response options. No misfit of items to the model. Box B: Test-retest, ICC = 0.70, sample size not reported. Box F: Hypotheses about convergent/divergent validity confirmed (correlation to BSA r = 0.5, less to non-related measures). Expected known group differences (according to BSA and CDAI groups) confirmed. Box I: Comparator is “clinically important change” based on PGA. Sparse information on properties of this measure, however it has high face validity (according to Cosmin). Nevertheless, it seems that										

the PGA of change addresses arthritis (not skin disease) in this paper (slightly different constructs). Significant differences in PSI scores between PGA "responders" vs "non-responders" were reported (approximately a 6 point difference).
Interpretation: N = 154 (in analysis except for test-retest where n was not reported). Missing data: 8% (excl.). Mean(SD) scores: 12.2(7.89) at baseline and 7.1(7.43) at week 12. Floor effect of individual items: 11-37% at baseline and 32.4-55.8% at week 12. Time frame test-retest: 2 weeks, responsiveness: 12 weeks. Ceiling effect for individual items: 4.5-7.1% at baseline and 1.9-2.6% at week 12.

PSD *Eng* [61] Excellent/+

Results: Box D: Thorough content validity evaluation during the development of the PSI instrument ensuring the comprehensiveness and relevance. Study population included between 34% (concept elicitation) and 50% (cognitive interview) with PsA .

NRS ITCH *Eng*[66] Fair/+

Results: Box D: Content validity confirmed by relevance to target population but the assessment of comprehensiveness less well described.
Interpretability: N = 22 PsA, 12 Psoriasis. Itching was a problem for 68% PsA patients and 100% of the psoriasis patients.

PAIN	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
VAS Pain (HAQ) *Eng*[28]								Poor/?	

Results: Box I: Responsiveness is tested in different ways but no evidence for responsiveness was achieved (small sample size of subanalysis/insufficient methods applied).
Interpretability: N= 70. Missing data: 12.5% (excl.) Mean(SD), Time 1: 0.97(0.72), Time 2: 0.83(0.81).

VAS Pain (HAQ) *Eng*[22] Fair/+

Results: Box F: Vague hypotheses and sparse information about measurement properties of comparators. Moderate/strong correlation between HAQ VAS pain score and tender points, function, stiffness and active joint count.
Interpretability: N = 99-114. Missing data: 13% (excl). Mean(SD), range score: 0.97(0.72). Floor/ceiling effect: NS

VAS pain *Eng*[43] MID

Interpretability: N = 200. Missing data: NS. Mean(SD) score (1^{st}/2^{nd} visit): 41.45(27.69)/38.65(28.84). Varying time interval (mean: 8.28 months between visits). MID(SD) estimates for improvement/worsening: -9.37(24.37)/13.96(22.05). Correlation between mean change in VAS pain and anchor (patient's rating of change): r_s = 0.448 (the authors did not aim to test responsiveness of VAS pain, only MID). Floor/ceiling effect: NS.

VAS pain *Chin*[50] Poor/? MID

Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in VAS pain: r_s=0.30.
Interpretability: N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Patients who continued treatment > 12 weeks: 67.8(23.3), patients treated <12 weeks: 62.7(16.8). MID for improvement: -14.71(31.25). MID for deterioration: -0.95(18.82). Effect size: -0.55, SRM: -0.49. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS

VAS pain *Norw*[45] MCII, PASS

Interpretability: Total number of PsA in the study: N = 1391. No. of PsA in the analysis: n = 847. Missing data: Reported for each part of the study. Mean(SD) score: 51.5(21.6) at baseline and 37.8(23.1) after 3 months. The anchoring questions were given after 3 months and asked about satisfaction (PASS) and improvement (MCII), respectively. PASS cut-points (ROC curves): 75% sensitivity cut-off: 38. 80% specificity cut-off: 30. AUC

0.80(95%CI 0.77-0.83). MCII cut-points (ROC curves): 75% sensitivity cut-off: -9.00 80% specificity cut-off: -18.0 AUC(95%CI): 0.76(0.73-0.79).

NRS pain *Eng*[44]	Poor/?	Poor/?	Poor/?	

Results: Box B: ICC (95%CI) = 0.83 (0.81-0.85) but number of patients in analysis (and % PsA) not specified. Box D: Not enough information to rate quality or results on content validity. Box F: No hypotheses or information about properties of comparators, PsA <50% of the population in most of the analyses. Correlation between patient-reported TJC and NRS pain reported for PsA (r=0.484) (n=57).
Interpretability: Total N = 462 (PsA 123, 26.6%). Number of PsA patients not reported for all analyses. Missing data: NS. Mean(SD)score (baseline): NS (1st assessment in reliability analysis: mean(SD) 6.4(1.2). Reliability time interval: 1 week. Floor/Ceiling effect: NS.

SF-36 BP *Eng*[26]	Poor/?		Fair/+	

Results: Box A**:** Cronbachα 0.90, unidimensionality not sufficiently checked. Box F: Convergent validity confirmed with 5/7 hypotheses fulfilled. Moderate-strong correlation with measures of function and disease activity. Known group validity showing significant difference to general population (no hypotheses about expected (magnitude of) differences were formulated).
Interpretability: N=113. Missing data: NS (all completed). Mean(SD) 61.5(2.47). Floor/ceiling effect: NS.

SF-36 BP *Eng*[28]	Poor/?		Poor/?	

Results: Box A: Unidimensionality not checked/reported. Cronbach α (0.80-0.91), no exact value was reported. Box I: Responsiveness is tested in different ways but no evidence for responsiveness was achieved (small sample size of subanalysis and or insufficient methods applied).
Interpretability: N=70, Missing data: 12.5% (excl.). Mean(SD) score at baseline/follow-up: 60.83.(23.99)/ 59.65(25.16). Floor/ceiling: NS.

SF-36 BP *Chin*[40]	Poor/?		Good/+	F/C

Results: Box A: Unidimensionality not sufficiently checked (No factor analysis). Cronbach α 0.838. Box F: Convergent validity (internal relationships): Scaling assumption (equal item variance, item-own scale, item-other scale) in consistency with hypotheses but for known group validity (external relationships) the hypotheses were vaguely stated.
Interpretability: N=168. Missing data: NS. Mean(SD) scale score: 48.54(21.85) Floor effect: 1.2%, Ceiling effect: 3.0%.

SF-36 BP *Chin*[50]			Poor/?	MID

Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in pain: r_s=-0.41.
Interpretability: N = 17-21 in analyses. Missing data: NS. Mean(SD)scores: a) Patients treated > 12 weeks: 30.7(12.5) and b) Patients treated <12 weeks: 42.8(21.4). Time frame (responsiveness): 52 weeks. MID for improvement: 12.35(19.83). MID for deterioration: -5.45(15.63). ES: 0.53, SRM:0.59. Floor/ceiling effect: NS

AIMS1 Pain *Eng*[21]			Fair/+	

Results: Box F: Sparse information about properties of comparators. However, AIMS pain correlated with measures of disease activity and function but not with measures of disease severity, as hypothesized.
Interpretability: N = 45. Missing data: NS. Mean(SD) scale score 2.1(1.7). Floor/ceiling effect: NS.

AIMS1 Pain *Eng*[25]			Poor/?	

Results: Box I: Not assessing correlation between change scores and no a priori hypotheses. AIMS1 at baseline and AIMS2 at follow up.
Interpretability: N = 65. Missing data: NS. Mean(SD) for Pain 3.08(1.99). Floor/ceiling effect: NS

AIMS1 Pain *Ita.*[27]						Fair/+			
Results: Box F: Sparse information about properties of comparators. Divergent validity (no correlation to clinical measures of disease severity) was confirmed and convergent validity (correlation to measures of function and disease activity) sufficiently confirmed. **Interpretability:** N=72. Missing data: NS. Means(SD) score 4.58(3.5), range 0-10. Floor/ceiling effect: NS									
AIMS2 Pain *Eng*[28]		Poor/?						Poor/?	
Results: Box A**:** Cronbach α (0.80-0.91) no exact value reported. No information on unidimensionality. Box I: Responsiveness is tested in different ways but evidence not achieved (inappropriate methods/too small sample size in subanalysis). Although 'patient's perceived change in health' (1 year) seems to have face validity (according to COSMIN) as a comparator, no hypotheses about expected magnitude of correlations or the expected SRM are stated. **Interpretability:** N= 70. Missing data: 12.5% (excl.) Mean(SD), range: Time 1: 3.90(2.78),0.00-9.50. Time 2: 3.69(2.85), 0.00-10.00.									
AIMS2 *Eng*[25]								Poor/?	
Results: Box I: Not assessing correlation between change scores and no a priori hypotheses. AIMS 1 is used at baseline and AIMS 2 after 4 years. **Interpretability:** N = 65. Missing data: NS. Mean(SD) for Pain 3.98(2.61). Floor/ceiling effect: NS									
AIMS2 Pain *Eng*[24]						Fair/+			
Results: A priori hypotheses about correlation with related measures of function, disease activity and disease activity sufficiently confirmed. (Moderate to high correlations with measures of function and disease activity (r = 0.34-0.56), but not with degree of joint deformity). **Interpretability:** N=124. Missing data: NS. Mean(SD) score 4.10(2.64). Floor/ceiling effect: NS.									
PATIENT GLOBAL	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Due to Psoriasis only									
NRS skin *Eng*[64]						Fair/+			F/C
Results: Box F: No hypotheses stated a priori. High correlations (>0.50) with related PROMS. Multivariable regression analyses reported that NRS skin was explained by skin problems, functional capacity, discomfort and pain (R^2 of model 0.806). **Interpretability**: N = 223. Missing data: <5%. Mean(SD) score: 4.1(3). Floor effect: ~22%. Ceiling effect: ~3%.									
VAS skin *Eng*[49]		Good/+				Poor/?			
Results: Box B: ICC(95% CI) = 0.78(0.72-0.83). Box F: No hypotheses or information about measurement properties of comparators. Multivariable regression with backward selection tested the influence of PASI, involvement of face, genitals, hands, buttocks and/or intergluteal and feet, psoriasis duration, sex, age, and occupation, The final regression model included PASI score and hand skin involvement, (R^2 = 0.35). Known group validity: No difference in VAS joint according to PsA phenotype. **Interpretability**: N = 319. Missing items: NS. Median(IQR) score: 30(11-60). Time frame (test-retest): 1 week. Floor/ceiling effect: NS									
Due to Arthritis only									
NRS joints *Eng*[64]						Fair/+			F/C
Results: Box F: No a priori hypotheses stated. Several correlations tested. High correlations (r >0.50) with related PROMs was found. Multivariable regression analyses reported that NRS-joints was well explained (R^2 of the model 0.778) by pain (β= 0.525), fatigue (β= 0.155), work and/or leisure									

activities (β = 0.178), depression (β = −0.104, P = 0.0193) and coping (β = 0.141).
Interpretability: N = 223. Missing data: <5%. Mean(SD) score: 5.6(2.5). Floor effect: ~7%. Ceiling effect: ~3%.

NRS joints *Eng*[44]		Poor/?	Poor/?

Results: Box D: Not enough information to rate quality or results on content validity. Box F: No hypotheses or information about properties of comparators, PsA <50% of the population in most of the analyses. Correlation between patient-reported TJC and NRS global reported for PsA (r=0.398) (n=57).
Interpretability: Total N = 462 (PsA 123, 26.6%). Number of PsA patients not reported for all analyses. Missing data: NS. Mean(SD)score (baseline): NS. Floor/Ceiling effect: NS.

VAS joints *Eng*[49]	Good/+	Poor/?	

Results: Box B: ICC(95% CI) = 0.86(0.81-0.89). Box F: No hypotheses or information about measurement properties of comparators. Multivariable regression with backward selection procedure to test the influence of TJC, dactylitis, enthestis, arthritis duration, sex, age, occupation. SJC: β(95%CI): 0.88 (0.24–1.52), TJC: 0.76 (0.47–1.06) and dactylitis: 9.45 (–0.10.18.99) were included in the final model. Known group validity: No difference in VAS Joints according to PsA arthritis phenotype (poly/oligo/mutilans/axial/distal/>1 type).
Interpretability: N = 319. Missing data: NS. Median(IQR) score: 47(22-69). Time frame (test-retest): 1 week. Floor/ceiling effect: NS.

Due to PsA

PGA by NRS *Eng*[64]		Fair/+	F/C

Results: Box F: Various correlations tested. No a priori hypotheses stated. High correlations (r >0.50) with related PROMs except skin PROMS (r = 0.33 (DLQI) and r= 0.52 (NRS embarrassment)). Multivariable regression analyses found PGA to be well explained by following (R^2 of the model 0.754): coping (β= 0.287), NRS pain (β= 0= 0.240), work and/or leisure activities (β = 0.141) and anxiety, fear and uncertainty (β = 0.109).
Interpretability: N = 223. Missing data: <5%. Mean(SD) score: 4.8(2.7). Floor effect: ~8%. Ceiling effect: ~3%.

PGA by NRS *Chin*[53]		Fair/+	

Results: Box F: Vague hypotheses. High/moderate correlation to related PROMs (NRS Pain: r_s =0.54, HAQ: r_s =0.54, SF-36 MCS: r_s = -0.47, SF-36 PCS: r_s =0.49, DAS28: r_s = 0.50), and less to clinician reported measures (all r_s <0.4). In multivariate regression analysis, PGA was associated with pain score, the PCS and MCS of the SF-36, and the PASI (these 4 variables explained 47.7% of the variance in PGA). Known group validity: Effect size for patients with different levels of disease severity ranged from 0.72 (social welfare dependence (y/n)) to -1.32 (fulfilment of MDA (y/n)).
Interpretability: N = 125. Missing data: None (patients were instructed to completion). Mean(SD) score: 4.56(2.32). Floor/ceiling effect: NS

PGA by VAS *Eng*[43]			MID

Interpretability: N = 200. Missing data: NS. Mean(SD) score (1^{st}/2^{nd} visit): 37.21(26.63)/35.24(27.96). Varying time interval (mean: 8.28 months between visits). MID(SD) estimates for improvement/worsening: -8.41(21.17)/11.53(21.03). Correlation between mean change in VAS global and anchor (patient's rating of change): $r_{s=}$ 0.490 (the authors did not aim to test responsiveness of VAS global, only MID) Floor/ceiling effect: NS.

PGA by VAS *Eng*[49]	Good/+	Poor/?	

Results: Box B: ICC(95% CI) = 0.87(0.83-0.90). Box F: No hypotheses or information about measurement properties of comparators. Multivariable regression with backward selection showed no impact of anxiety or depression on PGA. Final regression modal (R^2 = 0.73) showed that PGA was more

influenced by patient joint assessment (VAS joints): β(95%CI): 0.63(0.57–0.69) than Patient Skin Assessment (VAS skin): 0.30(0.27–0.37). I**nterpretability:** N = 319. Missing data: NS. Median(IRQ) score: 49(25-66). Time frame (test-retest): 1 week. Floor/ceiling effect: NS.		
PGA by VAS *Chin*[50]	Poor/?	MID
Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in VAS global: r_s=0.31. **Interpretability:** N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Treated > 12 / <12 weeks: 67.8(20.5)/58.2(16.6). MID (improve) /(deterioration) -11.76(25.06)/-2.86(18.88). Effect size: -0.50, SRM: -0.55. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS		
PGA by VAS *Norw*[45]		MCII, PASS
Interpretability: N = 1391. No. of PsA in the analysis: n = 847. Missing data: Reported for each study part. Mean(SD) score: 48.4(22.2) at baseline and 35.0(22.6) after 3 months. The anchoring questions were given after 3 months and asked about satisfaction (PASS) and improvement (MCII), respectively. PASS cut-points (ROC curves): 75% sensitivity cut-off: 35, 80% specificity cut-off: 25. AUC(95%CI): 0.78(0.75-0.81). MCII Cut-points (ROC curves): 75% sensitivity cut-off: -8.00, 80% specificity cut-off: -19.0 AUC(95%CI): 0.75(0.72-0.79).		
PGA by VAS *Ital*[63]	Fair/+	
Results: Box F: Correlations between PGA and different measures of disease activity reported but a priori hypotheses were very sparse. PGA had moderate to high correlation with composite disease activity measures and PhGA, and less correlation to unrelated measures like CRP. Good concordance between MDA and PGA<20 mm during follow up (kappa=0.72-0.74) Interpretability: N = 124 (minimum n = 75). Median(IQR) score at baseline: 59(45-70). Floor/ceiling effect: NS		

Table D continued

Identified PROMs listed according to Domain category	Reliability COSMIN BOX (A-C) Internal consistency (A)	Reliability (B)	Measure-ment error (C)	Validity COSMIN BOX (D-H) Content validity (D)	Structural validity (E)	Hypothe-ses testing (F)	Cross-cult. validity (G)	Criterion validity (H)	Responsiveness COSMIN BOX: I (I)	Relevant info on score in-terpre-tation
PHYSICAL FUNCTION	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
DFI *Chin*[34]	Good/-				Good/-	Poor/?				F/C
Results: Box A,E: Rasch: Item-trait chi^2 statistics: 35.7(df40, p=0.66). Person reliability: 0.85. Item separation: 3.83. Eight items showed misfit to the Rasch model. PCA: No evidence of a 2nd factor. Variance explained: 76%. DIF (sex) for 1 item and DIF (± sacroiliitis) for 1 item. Box F: No hypotheses or information about properties of comparators. Moderate/strong correlations with HAQ, PGA, pain (r= 0.44-0.76). **Interpretability:** N = 108. Missing data: NS. Mean(SD) score: 6.28(7.08). Floor effect: 31.1%.										
DASH *Eng*[29]						Good/-				
Results: Box E: Hypotheses not convincingly fulfilled (<75%): Correlation to measures of inflammatory joint count in upper extremity (R_s = 0.65), but not to measures of physical function (Grip strength, ACR functional class) or to measures of upper extremity damage. **Interpretability:** N = 50. Missing data: NS. Mean(SD) score: 27.5(24.6), Median(range) score:20.8(24-80.3). Floor/ceiling effect: NS										
BASFI *Chin*[34]	Good/+				Good/+	Poor/?				F/C
Results: Box A,E: Rasch: Item-trait chi^2 statistics: 27.8(df 20, p=0.11). Person reliability: 0.83. Item separation: 3.33. INFIT/OUTFIT values between 0.7-1.3. PCA: PCA: No evidence of a second factor. Variance explained: 78%. No DIF for sex, 1 item with DIF for ± sacroiliitis). Box F: No hypotheses or information about properties of comparators. Moderate/strong correlation with HAQ (r = 0.81), Pain (r = 0.52) and PGA (r = 0.49). **Interpretability:** N = 108. Missing data: NS. Mean(SD) score: 24.41(22.93). Floor effect 18.5%										
HAQ-DI *Eng*[22]						Fair/-				
Results: Box F: Vague hypotheses and sparse information about measurement properties of comparators. Less than 75% of hypotheses were fulfilled. Confirmed moderate/strong correlation to other measures of function (ACR functional class (r= -0.59(95%CI: -0.46 to -0.7)), Grip strength -0.63(-0.50 to 0.73)) and to measures of disease activity (Active joint count (r= 0.49(0.49 to 0.62)) and tender points (r = 0.54(0.40 to 0.66). Low correlation to other measures of disease activity/severity (effusion, stiffness, ERS, PASI) and to damage (ARA anatomic stage, damaged joint count). Moderate/strong correlation between HAQ VAS pain score and tender points, function, stiffness and active joint count. Multivariable regression identified 4 variables to influence on HAQ: Grip strength, ACR functional class, tender points and ESR. **Interpretability:** N = 99-114. Missing data: 13% (excl). Mean(SD),range score: 0.50(0.58), 0.00-2.00. Mean(SD) SpA vs not SpA: 0.61(0.64) vs 0.49(0.56)) (p=0.26) and for Fibromyalgia vs no Fibromyalgia: 1.32(0.49) vs 0.42(0.52). Floor/ceiling effect: NS										
HAQ-DI *Eng*[28]	Poor/?								Poor/?	
Results: Box A: Unidimensionality not checked. Cronbach-α = 0.80-0.91. Box I: Responsiveness is tested in different ways but no evidence for responsiveness was achieved (small sample size of subanalysis and or insufficient methods applied).										

Interpretability: N = 70. Missing data: 12% (excl.) Time frame (responsiveness): 12-18 months. Mean(SD) score: T1/T2: HAQ-DI 0.49(0.54)/ 0.46(0.58). Floor/ceiling effect: NS				
HAQ-DI *Eng*[33]	Good/+	Good/+		F/C
Results: Box A,E: Rasch: Good overall model fit. Person reliability: 0.75. Item separation: 2.06 logits. Misfit for the "Activity" item; INFIT MNSQ 1.58). **Interpretability:** N = 134 (PsA). Missing data: NS. Mean(SD) score: 0.5(0.59). Floor effect: 30.4%. (DIF reported for PsA vs RA)				
HAQ-DI *Eng*[43]				MID
Interpretability: N = 200. Missing data: NS. Mean(SD) scores: 1st visit: 0.732(0.677), 2nd visit 0.711(0.707). Varying time interval (mean: 8.28 months between visits). MID(SD) estimates for much improvement/much worsening: -0.362(0.432)/0.438(0.315). Correlation between mean change in HAQ and anchor (patient reported level of change): $r_{s=}$ 0.374 (the authors did not aim to test responsiveness of HAQ, only MID). Floor/ceiling effect: NS				
HAQ-DI *Eng*[51]				MID
Interpretability: N = 161. Missing data: NS. Mean(range)score: 1.16(0.13-2.88) at baseline. MID: 0.35. Minimal very important change: 0.45. Floor/Ceiling effect: NS				
HAQ-DI *Ital*[27]			Fair/-	
Results: No information about measurement properties of comparators. Hypotheses regarding high correlations (r>0.40) between HAQ scores and clinical measures of disease activity and disease severity were not sufficiently proven (less than 75%), the only moderate/strong correlations were between global HAQ and 1) duration of axial morning stiffness (r = 0.72) and 2) joint pain (r = 0.49). **Interpretability:** N = 72. Missing data: NS. Mean(SD) of the 8 area scores ranging between 0.82(0.79) (grip) to 1.15(0.95) (reach). Linearly transformed (0-100) global HAQ score mean(SD): 28.3(21.1). Floor/ceiling effect: NS				
HAQ-DI *Chin*[34]	Good/+	Good/+	Poor/?	F/C
Results: Box A,E: Rasch: Item-trait chi^2 statistics: 17.4(df 16, p=0.36). Person reliability 0.84. Item separation index 2.22. INFIT/OUTFIT values in the accepted interval (0.7-1.3) except for two items: 1) Dressing/grooming; OUTFIT 1.16. 2) Grip; OUTFIT 1.40, INFIT 1.41. HAQ limited by short item span (5.63 logits). PCA: No evidence of a second factor. Variance explained: 68%. DIF for item "Grip" according to sex. Box F: No hypotheses or information about properties of comparators. HAQ showed moderate/strong correlation to PGA (r_s = 0.54), Pain score (r_s = 0.56), TJC (r_s= 043) and BASFI ($r_{s=}$0.81), Dougados-FI (r_s=0.76) and SF-36 PF (r_s=0.80). **Interpretability:** N = 108. Missing data: NS. Mean(SD) score: 0.69(0.67). Floor effect: 24.5%.				
HAQ-DI *Hung*[42]			Fair/+	
Results: Box F: Sparse information about measurement properties of comparators. HAQ correlated to moderately/strongly to related measures: BASDAI (r_s = 0.59), PsAQoL (r_s = 0.64), PGA (r_s = 0.50), Pain score (r_s = 0.54). Known-group validity: Higher HAQ scores for patients with worse disease states. SRM (0.41-1.54). **Interpretability**: N = 183. Missing data: 6%.(excl.). Mean(SD) score: 1.0(0.7), median(range): 0.88 (0-3). Floor/ceiling effect: NS.				
HAQ-DI *Chin*[50]			Poor/?	MID
Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in HAQ: r_s=0.30. **Interpretability:** N = 17-21 in analyses. Missing data: NS. Mean(SD)scores: a) Patients treated > 12 weeks: 1.16(0.59) and b) Patients treated <12 weeks:				

1.02(0.68). Time frame (responsiveness): 52 weeks. MID for improvement: -0.27(0.06). MID for deterioration: 0.095(0.18). ES: -0.22, SRM: -0.22. Floor/ceiling effect: NS				
HAQ-DI *Thai.*[60]	Poor/?		Fair/+	F/C
Results: Box A: Unidimensionality not checked. Cronbach-α = 0.88. Box F: Sparse information about measurement properties of comparators. Hypothesis concerning strong correlation to BASDAI was fulfilled (r = 0.81). Moderate-strong correlation to other measures (PGA, pain, ERS, ASDAS). **Interpretability:** N = 47. Missing data: NS. Mean(SD) score: 0.47(0.47), median(range): 0.25(0-1.63). Floor effect: ~50%				
HAQ-S *Eng*[22]			Fair/-	
Results: Box F: Vague hypotheses and sparse information about measurement properties of comparators. High correlation (r>0.40) between the SpA scales (SPAR scales) of the HAQ-S and measures of spinal involvement (Finger-floor distance, chest expansion) was hypothesized but not sufficiently confirmed (r<0.25). No significant difference in HAQ-S scores between PsA patients with/without axial disease. **Interpretability:** N = 99-114. Missing data: 13% (excl.). Mean(SD)/range score: 0.53(0.57)/ 0.00-2.00. Mean(SD) scores for SpA vs non-SpA: 0.63(0.61) vs 0.43(0.51) and for Fibromyalgia vs not fibromyalgia: 1.30(0.50) vs 0.42(0.48). Floor/ceiling effect: NS				
HAQ-SK *Eng*[23]			Poor/?	
Results: Box F: No hypotheses and no information about measurement properties of comparators. Poor correlations (r<0.5) between original HAQ-DI and new HAQ skin scales as well as between the new HAQ-skin scales and PASI. **Interpretability:** N= 114. Missing data: 3%(excl). Mean(SD) score: HAQ: 0.55(0.60), HAQ-SK: 0.56(0.58). Skin-Scale: 0.60(0.77). Floor/ceiling effect:NS				
mHAQ *Norw*[45]				PASS,MCII
Interpretability: N = 1391. No. of PsA in the analysis: n = 845. Missing data: Reported for each part of the study. Mean(SD) scores: 1st visit: 0.63(0.44), after 3 months: 0.47(0.42). The anchoring questions were given after 3 months and asked about satisfaction (PASS) and improvement (MCII), respectively PASS Cut-points (ROC curves): 75% sensitivity cut-off: 0.50. 80% specificity cut-off: 0.14. AUC 0.75(95%CI 0.71-0.78). MCII Cut-points (ROC curves): 75% sensitivity cut-off: 0 80% specificity cut-off: -0.25. AUC(95%CI): 0.75(0.72-0.78).				
SF-36 PF *Eng*[33]	Good/+	Good/+		F/C
Results: Box A,E: Rasch: Good model fit, item separation 9.12 logits. No misfitting items. **Interpretability:** N = 134 (PsA). Missing data: NS. Mean(SD) score: 60.4(27.1). Floor effect 3.1%.				
SF-36 PF *Chin*[34]	Good/+	Good/+	Poor/?	F/C
Results: Box A,E: Rasch: item-trait chi^2 statistics: 24.3(df 20, p=0.23). Person reliability: 0.85. Item separation: 6.99. INFIT/OUTFIT values between 0.7-1.3. PCA: No 2nd factor. Variance explained: 89%. No DIF (gender or ± sacroiliitis). Box F: No hypotheses or information about properties of comparators. Moderate/strong correlation with HAQ (r= -0.80), PGA (r= -0.44) and VAS pain (r= -0.49). **Interpretability:** N = 108. Missing data: NS. Mean(SD) score: 63.33(25.5). Floor effect (Max score): 7.4%.				
SF-36 PF *Eng*[26]	Poor/?		Fair/+	
Results: Box A: Limited information on unidimensionality. Cronbachα 0.92. Box F: Convergent validity confirmed with 6/7 hypotheses fulfilled. Moderate-strong correlation with measures of function, disease activity and severity. Known group validity: Significant difference in scores compared to general population (but no hypotheses about expected magnitude of difference etc.)				

Interpretability: N=113. Missing data: NS (all completed) Mean(SD) scale score: 68.8(2.65). Floor/Ceiling effect: NS			
SF-36 PF *Eng*[28]	Poor/?		Poor/?
Results: BOX A: Unidimensionality not checked/reported. Cronbach α (0.80-0.91), no exact value was reported. Box I: Responsiveness is tested in different ways but no evidence for was achieved (small sample size of subanalysis/insufficient methods applied). **Interpretability**: N=70, Missing data: 12.5% (excl.). Mean(SD) score at baseline/follow-up: 70.07(25.63)/72.27(26.55). Floor/ceiling: NS.			
SF-36 PF *Chin*[40]	Poor/?	Good/+	
Results: Box A: Not sufficient reporting on unidimensionality. Cronbach α 0.913. Box F: Convergent validity (internal relationships): Scaling assumption (equal item variance, item-own scale, item-other scale) in consistency with hypotheses but for known group validity (external relationships) the hypotheses were vaguely stated. **Interpretability:** N=168. Missing data: NS. Mean(SD) scale score: 65.5(25.3). Floor effect: 1.8%, Ceiling effect: 7.7%.			
SF-36 PF *Chin*[50]		Poor/?	MID
Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in SF-36PF score: r_s= -0.34. **Interpretability:** N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Patients treated > 12 weeks: 51.5 (23.5) and 2) Patients treated <12 weeks: 57.7(21.8). MID for improvement: 4.41(14.99). MID for deterioration: -6.25(18.77). Effect size: 0.35, SRM: 0.37. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS			
SF-36 PCS *Chin*[40]	Good/+	Poor/?	
Results: Box E: Structural validity assessed by PCA, and a 2 factor model was supported, explaining 69.4% of the total variance. Box F: Only known group validity (general population vs. PsA) and no exact hypotheses stated a priori. **Interpretability:** N=168. Missing data: NS. Mean(SD) component summary score: 31.6(14.19)			
SF-36 PCS *Chin*[50]		Poor/?	MID
Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in SF-36 PCS: r_s= -0.43. **Interpretability:** N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Patients treated > 12 weeks: 22.3 (7.6) and 2) Patients treated <12 weeks: 28.0(9.3). MID for improvement: 3.74(8.51). MID for deterioration: -3.97(10.46). Effect size: 0.49, SRM: 0.55. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS			
CIAQ-FI *Eng*[44]	Poor/?	Poor/?	Poor/?
Results: Box B: Test retest Reliability ICC 0.912 (0.894-0.931) but number of patients (total and %PsA) in analysis not stated. Box: D: Not enough information available for rating the quality of content validity assessment or results. Box F: Sparse hypotheses and no information about properties of comparators. Most correlations reported for a mixed population (PsA<50%), for the PsA subset, the correlation between CIAQ-FI and HAQ was r = 0.927. The correlation between CIAQ-FI and PR-TJC: r=0.605 for the PsA subset (n=57). **Interpretability:** Total N = 462 (PsA 123, 26.6%). Number of PsA patients not reported for all analyses. Missing data: NS. Mean(SD) scores (baseline, PsA =26.6%) CASQ-F1 2.20 (0.7). Time frame test-retest: 1 week. Floor/ceiling effect: NS.			

Table D continued

Identified PROMs listed according to Domain category	Reliability COSMIN BOX (A-C)			Validity COSMIN BOX (D-H)					Responsiveness COSMIN BOX: I	Relevant info on score interpretation
	Internal consistency (A)	Reliability (B)	Measurement error (C)	Content validity (D)	Structural validity (E)	Hypotheses testing (F)	Cross-cult. validity (G)	Criterion validity (H)	(I)	
PHYSICAL FUNCTION	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
AIMS1 Mobility *Eng* [21]						Fair/-				
Results: Box H: Sparse information about properties of comparators. Less than 75% of hypotheses about correlation to measures of function, disease activity and severity were confirmed. **Interpretability:** N=145. Missing data: NS. Mean(SD) score: 0.5(1.1). Floor/ceiling effect: NS										
AIMS1 Physical *Eng*[21]						Fair/+				
Results: Box H: Sparse information about properties of comparators. Hypotheses about correlation to measures of function, disease activity and severity were sufficiently confirmed. **Interpretability:** N=145. Missing data: NS. Mean(SD) score: 3.8(1.5). Floor/ceiling effect: NS										
AIMS1 Dexterity *Eng*[21]						Fair/+				
Results: Box H: Sparse information about properties of comparators. Hypotheses about correlation to measures of function, disease activity and severity were sufficiently confirmed. **Interpretability:** N=145. Missing data: NS. Mean(SD) score: 2.6(1.7). Floor/ceiling effect: NS										
AIMS1 Household *Eng*[21]						Fair/+				
Results: Box H: Sparse information about properties of comparators. Hypotheses about correlation to measures of function, disease activity and severity were sufficiently confirmed. **Interpretability:** N=145. Missing data: NS. Mean(SD) score: 0.9(0.7). Floor/ceiling effect: NS										
AIMS1 ADL *Eng*[21]						Fair/-				
Results: Box H: Sparse information about properties of comparators. Less than 75% of hypotheses about correlation to measures of function, disease activity and severity were confirmed. **Interpretability:** N=145. Missing data: NS. Mean(SD) score: 0.3(0.7). Floor/ceiling effect: NS										
AIMS1 Physical Component *Eng*[25]									Poor/?	
Results: Box I: No hypotheses and different versions of AIMS used at baseline and follow-up (AIMS 1 and AIMS 2), no correlation between AIMS change scores and change scores of clinical measures reported. **Interpretability**: N= 65. Missing data: NS. Mean(SD) score Physical: 1.47(1.64). Floor/Ceiling. NS										
AIMS1 Physical *Ital*[27]						Fair/-				
Results: Sparse information on properties of comparators. Seems that hypotheses only concerned 1 of the function scales (the physical activity scale) and hypotheses were not convincingly fulfilled (less than 75%). A "close" correlation between AIMS physical function scale and disease activity measures was										

expected but the majority of correlations presented was with Pearson's r <0.4. Interpretability: N=72. Missing data: NS. mean(SD) score: 5.97(3.1). Floor/ceiling effect: NS	
AIMS2 Physical Component[25]	Poor/?
Results: Box I: No hypotheses and different versions of AIMS used at baseline and follow-up (AIMS 1 and AIMS 2), no correlation between AIMS change scores and change scores of clinical measures reported. **Interpretability**: N= 65. Missing data: NS. Mean(SD) score Physical: 1.37(1.36)Floor/Ceiling. NS	
AIMS2 Physical Component *Eng*[28]	Poor/?
Results: Box I: Responsiveness is tested in different ways but no evidence for responsiveness was achieved (small sample size of subanalysis and or insufficient methods applied). **Interpretability**: N= 70. Missing data: 12.5% (excl.) Mean(SD), range: Baseline: 1.04(1.20), 0.00-5.42, follow-up: 1.29(1.53), 0.00-5.35.	
AIMS2 Mobility *Eng*[24]	Fair/+
Results: Box F: Hypotheses about correlation to measures of disease activity and function sufficiently fulfilled. **Interpretability:** N=124. Missing data: NS. Mean(SD) scores 1.21(1.68), range 0-7. Floor/ceiling: NS	
AIMS2 Physical *Eng*[24]	Fair/+
Results: Box F: Hypotheses about correlation to measures of disease activity and function sufficiently fulfilled. **Interpretability:** N=124. Missing data: NS. Mean(SD) scores 3.04(2.77), range 0-10. Floor/ceiling: NS	
AIMS2 Dexterity *Eng*[24]	Fair/+
Results: Box F: Hypotheses about correlation to measures of disease activity and function sufficiently fulfilled. **Interpretability:** N=124. Missing data: NS. Mean(SD) scores 1.58(1.97), range 0-8.5. Floor/ceiling: NS	
AIMS2 Selfcare *Eng*[24]	Fair/-
Results: Box F: Hypotheses about correlation to measures of disease activity, disease severity and function not fulfilled. **Interpretability:** N=124. Missing data: NS. Mean(SD) 0.60(1.90), 0-10	
AIMS2 Household *Eng*[24]	Fair/-
Results: Box F: Hypotheses about correlation to measures of disease activity, severity and function sufficiently fulfilled. **Interpretability:** N=124. Missing data: NS. Mean(SD) scores 0.52(1.05), range 0-6.25. Floor/ceiling: NS	
AIMS2 Arm Function *Eng*[24]	Fair/+
Results: Box F: Hypotheses about correlation to measures of disease activity and function sufficiently fulfilled. **Interpretability:** N=124. Missing data: NS. Mean(SD) scores 0.70(1.20), range 0-5.5. Floor/ceiling: NS	

Identified PROMs listed according to Domain category	Reliability COSMIN BOX (A-C)			Validity COSMIN BOX (D-H)					Responsiveness COSMIN BOX: I	Relevant info on score interpretation
	Internal consistency (A)	Reliability (B)	Measurement error I	Content validity (D)	Structural validity I	Hypotheses testing (F)	Cross-cult. Validity (G)	Criterion validity (H)	(I)	
HRQoL/Disease impact	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
PsAQoL *Eng*[30]	Good/+	Fair/?		Excellent/+	Good/+	Fair/+				
Results: Box A, E: Unidimensionality checked, Cronbach-α = 0.91. Rasch: Item trait interaction Chi^2: 96.1 (df=80; p=0.106), overall item fit mean(SD): 0.183(1.115), person fit mean(SD): 20.232(0.807). Person separation index of 0.922. Box B: Reliability only reported by Spearman correlation (not considering systemic error). Box D: PsAQoL was developed by qualitative interviews with PsA patients and modified to obtain relevance, comprehensiveness and interpretability/feasibility. Box F: Vague hypotheses. Statistical methods used for correlation analyses were not reported, but moderate/high correlations to related measures (Nottingham Health Profile scale scores (r = 0.52-0.75), overall health VAS (r = 0.64) and quality of life VAS (r = 0.65)) were reported. Known group validity: Significant differences in PsAQoL scores according to disease severity level and perceived current health status. **Interpretability:** N = 211-286. Missing data: Excluded (8% for 2nd survey). Median(IRQ);range score: 9(5-13);0-20. Time frame (test-retest): approximately 2 weeks (not clearly stated). Floor/ceiling effect: NS.										
PsAQoL *Eng*[35]						Poor/?			Poor/?	
Results: Box F: No hypotheses and small sample size (<30). Moderate/high baseline correlation to HAQ (r_s = 0.69), PGA (r_s = 0.44) and low correlation to (PhGA, TJC, SJC, DAS28 and PASI (all r_s < 0.3). The correlation between PsAqoL and other measures were more pronounced at follow up visits (e.g., at 3 months, correlations to HAQ, PGA, DAS28, PhGA were all r_s >0.46). Box I: Small sample and no hypotheses. No correlations between change scores reported. SRM at 3 months: 0.71 and at 6 months: 0.41. **Interpretability:** N = 28. Missing data: NS. Mean(SD) scores: 13.46(5.15) at baseline, 10.67(6.32) after 3 months and 10.5(6.92) after 6 months. Time frame (responsiveness): 6 months. Floor/ceiling effect: NS.										
PsAQoL *Eng/Chin[71]*	Poor/?	Fair/+		Poor/?		Fair/+	Fair			
Results: No information on unidimensionality, Cronbach alpha = 0.92. Box B: Test-retest reliability, ICC = 0.92. Box D: Too sparse information on content validity in the result section (e.g., unknown comprehensiveness). Box F: Vague hypotheses. Moderate-high correlation between PsAQoL and pain, PGA, PhGA, SF-36 subscales and summary scores, CPDAI. Known group proven by greater PsAQoL scores in patients poorer physical health, higher disease activity state (CPDAI, MDA). Box G: No cross-cultural validation performed, translation described. **Interpretability:** N=98 (67% Eng, 33% Chin), Missing data: NS. Mean(SD) scores at baseline: 4.5(5.2). Time frame (reliability) 2 weeks. Floor/Ceiling: NS										

COSMIN BOX	A	B	C	D	E	F	G	H	I	
PsAQoL *Swe*[41]	Good/+	Poor/?		Good/+		Fair/+	Fair			F/C
Results: Box A: Cronbach-α = 0.91 and authors refer to another study reporting on unidimensionality. Box B: No reliability results reported. Box D: Comprehensiveness and relevance assessed and confirmed. Box F: Sparse information about information about measurement properties of comparators. Moderate/high correlation to related measures (NHP scales: r_s = 0.53-0.80, NHPD: r_s =0.87) and known group validity according to PGA and flare-status, all in accordance with hypotheses. Box G: No cross-cultural validity assessment only translation. Only forward translation. **Interpretability**: N = 123. Missing data: 6-33% missing responses (excl.). Mean(SD) scores: 5.8(5.2). Floor effect: 19%, Ceiling effect 0%.										
PsAQoL *Hung*[42]						Fair/+				
Results: Box F: Sparse a priori hypotheses and information about measurement properties of comparators. Moderate to high correlations to HAQ (r_s = 0.64), BASDAI (r_s=0.62), PGA (r_s=0.52), VAS pain (r_s=0.54). Known group validity: Higher scores for patients with more severe disease level, SRM (0.53 to 1.70). **Interpretability:** N = 183. Missing data: 3% (excl.). Mean(SD) score: 7.7(6.0). Median(range) score: 7.0(0-20). Floor/ceiling effect: NS.										
PsAQoL *Dutch*[55]	Poor/?	Good/+	Good/?	Good/-		Fair/+	Fair			
Results: Box A: Unidimensionality not checked. Cronbach α = 0.92. Box B: r_s = 0.89 (95%CI 0.85-0.92) and Bland-Altman analysis demonstrating no systematic error between the administration. Box C: LoA between -5.3 and 5.1 (out of 20) but MIC not defined. Box D: 50% of patients reported that items were missing. Box F: Correlation to (somehow) related measures as expected (HAQ (r_s = 0.72), Skin-17 Psychosocial scale (r_s =0.40) and Skin-17 Symptom scale (r_s=0.46)). Known group validity: Higher scores for patients with worse PGA and higher disease activity. Box G: Only translation, no cross-cultural validity assessment. Only forward translation. **Interpretability**: N = 211 (134 for test-retest, 175 for internal consistency, 156 for convergent validity). Missing data: Reported (excl.) Median(range) score: 5.00(0-20). Time frame (test-retest): 2 weeks. Floor/ceiling effect: NS										
AIMS global score *Ita.*[27]						Poor/?				
Results: Box F: No hypotheses for the global AIMS scale. Sparse information about measurement properties of comparators AIMS global score was related to various measures of function and disease activity, the only strong correlations found were between AIMS global score and 1) morning stiffness of axial joints (r=0.63) and VAS pain (r=0.64). **Interpretability:** N = 72. Missing data: NS, Mean(SD) scale scores : NS. Floor/ceiling effect: NS.										
PSAID-9 *Eng*[58]	***note***	Good/+				Fair/+	Good		Poor/?	PASS,F/C
Results: Box A: : Cronbach-α = 0.93 but unidimensionality not reported. According to the authors PsAID is based on a formative model and therefore the internal consistency is not rated. Box B: Test –retest ICC (95%CI) = 0.94(0.91-0.96). Box F: Hypotheses were vaguely stated. High/moderate correlations (r_s= 0.408-0.845) with related measures (PGA, pain, HAQ, DLQI, SF-36 component summary scores, EQ-5D, DAS28). Box G: No cross-cultural validation, only translation. Box I: No hypotheses or change score correlations provided. Patients with self-reported improvement were included in the analyses and SRM was 0.90 (95% CI 0.88 to 0.92). **Interpretability**: N = 439 (in the validation part). Missing data: 1% (excl.). Mean scores: NS. PSAID PASS cut-off: 4. Time frame for test-retest: 2–10 days, and for responsiveness: 10–16 weeks.										

COSMIN BOX	A	B	C	D	E	F	G	H	I	
PsAID-12 *Eng*[58]	***note***	Good/+				Fair/+	Good		Poor/?	PASS,F/C
Results: Box A: Cronbach-α = 0.94 but unidimensionality not checked. According to the authors PsAID is based on a formative model and therefore the internal consistency is not rated. Box B: ICC(95%CI) = 0.95(0.92-0.96). Box F: Hypotheses were vaguely stated. High/moderate correlations (r_s = 0.422-0.843) with related measures (PGA, pain, HAQ, DLQI, SF-36 component summary scores, EQ-5D, DAS28). Box G: No cross-cultural validation, only translation. Box I: No hypotheses about the expected SRM for the correlations or change score correlations provided. In patients who reported improvement after treatment, the SRM was 0.91 (95% CI 0.89 to 0.93). **Interpretability**: N = 439 (in the validation part). Missing data: 1% (excl.). Mean scores: NS. PSAID Patient Acceptable symptom state cut-off: 4. Time frame for test-retest: 2–10 days, and for responsiveness: 10–16 weeks. Floor/ceiling effect: <1%										
PsAID-9/12[68]					Excellent/+					
Results: This paper is an elaboration of the PsAID development paper by Gossec et al (above). A further description of the involvement of patient partners in the development of PsAID is given providing strong evidence for content validity of the questionnaire.										
PsAID-12$_{touch}$ *Ital[72]*						Fair/+	Fair/+		MDA cut-off	
Results: Box F: Hypotheses and (psychometric) information on all comparators not thoroughly described. Convergent and know group validity examined, and expected relations stated a priori. PsAID-12 touch version correlated acceptably with PASDAS, DAPSA, HAQ and PhGA (r_s= 0.63-0.67), and the ability of PsAID-12 touch to discriminate between known groups (disease activity) was comparable to other measures of disease activity and function with ROC AUC (95% CI) =0.937 (0.898-0.975). Box H: The touch version of PsAID was compared to the original paper PsAID version (gold standard) and ICC for items were all >0.80. Mean difference (limit of agreement): 022(-0.60 to 1.04) by Bland Altman plot. For both boxes, information on handling of missing items/data was not clearly reported. **Interpretability:** N=159. Missing data: NS. Median(SD) scores of paper vs touch version: 3.60(1.96-4.78) vs 3.17(1.93-4.54). PsAID-12 touch version cut-off value for MDA: 2.5. Floor/ceiling effect: NS. Mean(SD) time of completing touch vs paper version: 1.7(2.21) vs 2.25(2.88).										
PsAID-12 *Ital[73].*	*Note*				*Note*	Fair/+			Cut-offs	
Results: Box F: Sparse hypotheses, convergent and known-group validity confirmed as PsAID-12 correlated well with cDAPSA, DAPSA, DLQI, PGA (r_s 0.489-0.867), and PsAID scores were increased in groups with higher compared to lower disease activity measured by cDAPSA. Box A, E: Internal consistency and structural validity were assessed in the study (the factor analysis found a 2-factor structure of PsAID ("symptoms" and "skin")) but these properties were not rated because PsAID is based on a formative rather than reflective model. **Interpretability:** N=144. Missing data: NS. PsAID median scores in categories defined by cDAPSA disease state: Remission (REM): 0.5, Low disease activity (LDA) 2.6, Moderate disease activity (MoDA); 6.2, High disease activity (HAD): 7.3. Cut-off values defined: REM≤1.4, LDA (>1.4 to ≤4.1), MoDA (> 4.1 to ≤6.7), HAD (>6.7).										
PAIP *Ital*[56]						Poor/?				
Results: Box F: No hypotheses stated a priori, sparse information on comparators. Moderate/high correlation (r>0.5) between PAIP subscales and presumably related measures (MOS-SF-36 subscales, McGill Pain Questionnaire subscales, Zeung Self-rating depression/anxiety scales).										

Interpretability: N = 123 (PsA: n = 82). Missing data: NS. Mean(SD) scores: NS. Floor/ceiling effect: NS. Floor/ceiling effect: <1%							
VITACORA-19 *Span*[59]	Fair/+	Good/+	Good/+	Fair/+	Fair/+	Poor/?	MCID,F/C
Results: Box A,E: Unidimensionality checked. Cronbach-α = 0.95. PCA: 1 factor explaining 55.8% of the observed variance. Box B: ICC= 0.94. Box D: Relevance and comprehensiveness assessed. Box F: Sparse information about measurement properties of comparators and vague hypotheses. Moderate/high correlation to EQ-5D VAS (r = 0.493), PhGA (r = 0.566), BSA (r=-0.664), DAS28 (r = 0.423). Known group validity: Differences in scores between PsA patients and healthy controls. Box I: No hypotheses, poor description of comparators and their measurement properties and no exact results provided (only reporting correlation between change scores: r<0.7). Effect size (0.2-0.8) for patients who experienced at least a small improvement in global health from 0-6 months. **Interpretability**: N = 209 PsA (n = 97 in test-retest). Missing data: Provided for each analysis (excl.). Mean(SD) score: 56.24(24.8). MCID: 8 point. Time frame test-retest: 10 days, responsiveness: 6 months. Floor, Ceiling effects: <1%.							
VITACORA-19 *Turk.[69]*	Poor/?	Fair/+	Poor/?	Fair/+	Fair/		
Results: Box A: Factor analysis performed but sample size insufficient (n=61). Cronbach-α = 0.96. Box B: Test-retest with ICC reported for each item (0.77-0.98). No evidence that patients were stable in the interim period and no description of missing data. Box E: Sample too small for factor analysis. Box F: Hypotheses were vague but correlation to HAQ (-0.60), Nottingham Health Profile items (-0.54 to -0.72) and to VAS Pain (-0.43). Box G: Translation, no cross-cultural validation only translation. **Interpretability:** N=61. Missing data: NS. Mean(SD) scores: 66.9(20.2). Time frame test-retest: 10-15 days. Floor/Ceiling effect: NS							
PsoDisk *Ital*[62]							Poor/?
Results: Box I: Small sample size, sparse hypotheses and no information about measurement properties of comparators. High correlation (r = 0.97) to PASI (measure of skin disease activity) **Interpretability:** N = 19. Missing data: NS. Mean(SD) scores at baseline ranging from 4.10(3.40) (for sleep) to 7.13(3.35) for skin involvement. PsoDisk change scores for each items reported at different time points showing significant difference from baseline. Time frame (responsiveness): 48 weeks. Floor/ceiling effect: NS.							
CIAQ-QoL *Eng*[44]		Poor/?	Poor/?	Poor/?			
Results: Box B: Test-retest: Number and % PsA patients in analysis was not specified. ICC(95%CI) 0.912(0.894-0.931). Box D: Not enough information to score quality or result regarding content validity. Box F: Sparse hypotheses and no information about measurement properties of comparators and not stated how many PsA was included, except for correlation between CIAQ-QoL and PR-TJC (r=0.08) for PsA subset (n=57). **Interpretability:** Total N = 462 (PsA 123, 26.6%). Missing data: NS. Number of PsA patients not reported for all analyses. Mean(SD) baseline score: 2.13 (0.9). Test-retest time interval: 1 week. Floor/ceiling effect: NS.							

Table D continued

Identified PROMs listed according to Domain category	Reliability COSMIN BOX (A-C)			Validity COSMIN BOX (D-H)					Responsiveness COSMIN BOX: I	Relevant info on score interpretation
	Internal consistency (A)	Reliability (B)	Measurement error (C)	Content validity (D)	Structural validity (E)	Hypotheses testing (F)	Cross-cult. validity (G)	Criterion validity (H)	(I)	
HRQoL/MULTIDIM.	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
WTP *Eng*[46]				Poor/?		Fair/+				
Results: (Pilot study). Box D: Content validity of the tool was reviewed by rheumatologists not PsA patients during the development phase. The proportion of patients confirming an impact of PsA on these domains varied between 35%-88%. Box F: Correlation between median WTP amounts were higher between related than non-related domains. However, one of the 4 domains that patients ranked as most impacted by PsA was associated with a lower WTP amount than some domains ranked at less impacted by PsA. **Interpretability:** N = 60. Missing data: NS. Median(IQR) scores of WTP for relief of the 8 domains: Lowest WTP amount (concentration): 7500(1000-50,000), highest WTP amount (Physical comfort, Sleep, work): 10,000(5000-75,000).										
IPBOD *Eng*[74]	Poor/?			Poor/?		Poor/?				
Results: Box A: Unidimensionality not assessed and small sample size. IPBOD may be based on a formative rather than reflective model why the assessment of internal consistency may be irrelevant. Box D: Not enough information available. Box F: Sparse hypotheses and information on comparator properties. Small sample size. Moderate correlation between total IPBOD and DLQI scores (r_s = 0.65) and similar correlation between "subscales" of IPBOD and related dimensions of DLQI **Interpretability:** N=16. Missing data: 6% (excl.). Mean(SD) score: 4.9. Floor/ceiling effects: NS.										
FATIGUE	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
FACIT-Fatigue *Eng*[32]	Poor/?	Fair/+				Fair/+				
Results: Box A: Unidimensionality not checked. Cronbach-α = 0.96. Box B: ICC = 0.95. Box F: Vague hypotheses. High correlation to related measures (mFSS score r= -0.79) and less to unrelated measures. Known group validity: Differences in scores between patients with vs. without overwhelming fatigue and between patients with PsA vs. general population). **Interpretability**: N = 135 (test retest: n = 73). Missing data: NS. Mean(SD) score: 35.8(12.4). Time frame (test-retest): 1 week. Floor/ceiling effect: NS.										
NRS fatigue *Eng*[38]						Poor/?			Poor/?	
Results: Box F: No a priori hypotheses. Correlation with related measures (GH, VAS pain, HAQ: r= 0.47-0.54). Regression analysis: At baseline and (3 months) GH, HAQ and pain explained 29% (38%); 23% (30%); and 22% (28%) of the variance in NRS fatigue scores, respectively. Box I: SRM and Effect size for the fatigue score and other measures shown only by box plots (no exact values provided). Correlation with other change scores not assessed and no hypotheses formulated.										

Interpretability: N = 41. Missing data: NS. Mean(SD) scores: 5.71(2.32) at baseline and 3.96(2.06) after 3 months. Time frame (responsiveness): 3 months. Floor/ceiling effect: NS.										
NRS fatigue *Eng*[44]			Poor/?		Poor/?				Poor/?	
Results: Box B: ICC (95%CI) = 0.85(0.83-0.87) but not stated how many patients (and % PsA) included. Box D: Not enough information to rate quality or results on content validity. Box F: No hypotheses or information about properties of comparators, PsA < 50% of the population in most of the analyses. Correlation between patient-reported TJC and Fatigue reported for PsA (r=0.447) (n=57). **Interpretability**: Total N = 462 (PsA 123, 26.6%). Number of PsA patients not reported for all analyses. Missing data: NS. Mean(SD)score (baseline): NS (but for 1st reliability assessment the mean(SD) score was: 7.6(0.47) (% PsA not clear). Floor/Ceiling effect: NS.										
VAS fatigue *Eng*[43]										MID
Interpretability: N = 200. Missing data: NS. Mean(SD) scores: (1st/2nd visit): 40.82(31.68)/38.30(30.42). Varying time interval (mean: 8.28 months between visits). MID(SD) estimates for improvement/worsening: -8.15(23.52)/3.63(27.615). Correlation between mean change in VAS fatigue and anchor (patient's rating of change): $r_{s=}$ 0.239 (the authors did not aim to test responsiveness of VAS fatigue, only MID) Floor/Ceiling: NS.										
SF-36 VT *Eng*[26]			Poor/?						Fair/-	
Results: Box A: Cronbachα 0.90, insufficient information on unidimensionality. Box F**:** Less than 75% of hypotheses about convergent validity (moderate correlation to measures of function, disease activity and severity) were confirmed. No statistically significant difference in scores compared to general population (known group validity). **Interpretability**: N=113. Missing data: NS. Mean(SD) scale score: 57.5(2.52). Floor/ceiling. NS-										
SF-36 VT *Chin*[40]			Poor(?)						Good(+)	F/C
Results: Box A: Unidimensionality not sufficiently checked. Cronbach alpha = 0.83. Box F: Internal convergent validity hypotheses confirmed. Known group validity (external relationships): No a priori hypotheses. **Interpretability**: N= 168. Missing data: NS (all completed). Mean(SD) scale score: 50.42(22.01). Floor/Ceiling effect: 0.6/2.0.										
SF-36 VT *Chin* [50]									Poor/?	MID
Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in vitality score: r_s= -0.28. **Interpretability:** N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Patients treated > 12 weeks:43.9 (9.3) and 2) Patients treated <12 weeks: 41.8(11.2). MID for improvement: 7.94(11.46). MID for deterioration: -5.25(15.68). Effect size: 0.28, SRM: 0.35. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS										
PARTICIPATION	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
SRPQ *(IM, SR,ST) Eng[52]*	Poor/?	Fair/+	Fair/?	Fair/+		Fair/+				MDC
Results: Box A: Unidimensionality not checked. SRPQ-Role importance scale: Cronbach α = 0.82, inter-item correlations: 0.09-0.75. SRPQ-Satisfaction (time spent and role): Cronbach α >0.93, inter-item correlations: 0.36-0.89. Box B: PsA group: Role importance: ICC (95%CI) = 0.79(0.60-0.90), Satisfaction with time spent: ICC (95%CI) = 0.94(0.88-0.97), Satisfaction with role performance: ICC(95%CI) = 0.96(0.92 to 0.98). Box C: No MIC defined. Box D: Sparse information. SRPQ was evaluated by cognitive debriefing by 15 patients but proportion with PsA was not reported. Box F: Hypotheses generally fulfilled, correlation with related measures of participation: r = 0.66-0.68.										

Interpretability: N = 109 (PsA=65). Missing data: NS. Mean(SD) summary scores (PsA): Importance scale: 3.81(0.48), satisfaction with time spent scale: 3.47(0.78) and satisfaction with role performance scale: 3.44(0.87). MDC (PsA): 1) Role importance scale: 0.86, 2) Satisfaction with spent scale: 0.75 and 3) Satisfaction with the role performance scale: 0.68. Time frame (test-retest): 2-3 weeks. Floor/ceiling effect: NS.			
WPS (all subscales)	Fair/+	Fair/+	F/C
Eng[57]			
Results: Box F: Hypotheses about divergent validity with low correlations (r<0.4) to unrelated measures (DAS28, CRP, PASI, HAQ-DI, SF-36 MCS, PCS, PsAQoL, EQ-5D, DLQI) were fulfilled. Known group validity: Differences in WPS scores between groups with different level of HRQoL (PsAQoL, SF-36), disease activity (DAS28, PASI) and disability (HAQ) confirmed. Box I: Hypotheses about known group validity stated a priori and confirmed with significantly greater improvements in household and patient workplace productivity observed in ACR20 responders versus nonresponders at week 12 except for item/subscale 2 and 8. **Interpretability:** N = 409. Missing data max 41% for a WPS item (not all items relevant for every patient). Missing data among patients expected to answer an item max. 0.7%. Mean(SD) scores reported for item 2-9. Time frame (responsiveness): 12 weeks. Floor effect (items): Max. 77%, Ceiling effect: Max 6.9%. ES for changes in work productivity for ACR20- or HAQ-DI MCID responders were small to moderate.			
AIMS1 Social activity *Eng*[21]	Fair/?		
Result: Sparse information about properties of comparators, and all comparators represent unrelated constructs (function, disease activitity or severity) why no results score is generated **Interpretability:** N=145. Missing data: NS Mean(SD) scale score: 2.6(1.6). Floor/ceiling effect: NS			
AIMS2 Social activity *Eng*[24]	Fair/?		
Result: Sparse information about properties of comparators and all comparators represent unrelated constructs (function, disease activitity or severity), why no results score is generated**.** **Interpretability: Interpretability:** N = 124. Missing data: NS. Mean(SD) scale score: 4.79(2.09). Floor/ceiling effect: NS			
AIMS2 Work *Eng*[24]	Fair/?		
Result: Sparse information about properties of comparators and all comparators represent unrelated constructs (function, disease activitity or severity) why no result is generated. **Interpretability:** N=124. Missing data: 51% (this scale is only relevant for employed patients). Mean(SD)scale score:1.41(1.80). Floor/ceiling effect: NS**:**			
AIMS2 Social component *Eng*[28]	Poor/?		
Results: Box I: Responsiveness is tested in different ways but no evidence for responsiveness was achieved (small sample size of subanalysis and or insufficient methods applied). **Interpretability:** N = 80. Missing data: 12% (excl.). Mean(SD) for scale scores at $1^{st}/2^{nd}$ visit: 3.33(1.86)/3.44(1.79). Time interval 12-18 months. Floor/ceiling effect: NS			
SF-36 RE *Eng*[26]	Poor/?	Fair/?	
Results: Box A: Limited information on unidimensionality. Cronbachα 0.92. Box F: Sparse information on properties of comparators, and only convergent validity assessed by comparing to unrelated measures (function, disease activity, disease severity. Hypotheses about significant difference			

between scores of PsA and general population confirmed, but no hypotheses about expected magnitude, not sufficient to generate a positive score for box F.
Interpretability: N=113. Missing data: NS (all completed) Mean(SD) scale score: 71.4 (4.44)Floor/Ceiling effect: NS

SF-36 RE *Chin*[40]	Poor/?	Good/+	F/C

Results: Box A: Unidimensionality not sufficiently checked. Cronbach α 0.868. Box F: Convergent validity (internal relationships): Scaling assumption (equal item variance, item-own scale, item-other scale) in consistency with hypotheses but for known group validity (external relationships) the hypotheses were vaguely stated.
Interpretability: N=168. Missing data: NS. Mean(SD) scale score: 48.41(44.53) Floor effect: 39.3%, Ceiling effect: 36.9%.

SF-36 RE *Chin*[50]	Poor/?	MID

Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in SF-36 RE score: r_s= -0.23.
Interpretability: N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Patients treated > 12 weeks: 19.4(34.9) and 2) Patients treated <12 weeks: 20.5(40.0). MID for improvement: 3.96(52.60). MID for deterioration: -10.0(37.62). Effect size: 0.27, SRM: 0.26. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS

SF-36 RP *Eng*[26]	Fair/-

Results: Box A: Limited information on unidimensionality. Cronbachα 0.92. Box F: Sparse information on properties of comparators, hypotheses about correlation to related measures (function, disease activity, disease severity) not sufficiently confirmed (3/7). Statistically significant difference in scores compared to general population (as hypothesized) but not enough to generate a positive score for box F.
Interpretability: N=113. Missing data: NS (all completed) Mean(SD) scale score: 65.8(4.30). Floor/Ceiling effect: NS

SF-36 RP *Chin*[40]	Poor/?	Good/+	F/C

Results: Box A: Unidimensionality not sufficiently checked. Cronbach α 0.888. Box F: Convergent validity (internal relationships): Scaling assumption (equal item variance, item-own scale, item-other scale) in consistency with hypotheses but for known group validity (external relationships) the hypotheses were vaguely stated.
Interpretability: N=168. Missing data: NS. Mean(SD) scale score: 41.07(42.41) Floor effect: 42.9%, Ceiling effect: 26.8%.

SF-36 RP *Chin*[50]	Poor/?	MID

Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in SF-36 RP score: r_s= -0.29.
Interpretability: N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Patients treated > 12 weeks: 19.4(34.9) and 2) Patients treated <12 weeks: 20.5(40.0). MID for improvement: 11.76(44.30). MID for deterioration: -11.25(36.70). Effect size: 0.41, SRM: 0.39. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS

SF-36 SF *Eng*[26]	Poor/?	Fair/?

Results: Box A: Limited information on unidimensionality. Cronbachα 0.88. Box F: Sparse information on properties of comparators, convergent validity assessed by comparing to unrelated clinical measures (function, disease activity, disease severity). Known group validity: Significant difference in scores compared to the general population (as hypothesized) but insufficient information to generate a positive score for box F.
Interpretability: N=113. Missing data: NS (all completed) Mean(SD) scale score: 65.8(4.30). Floor/Ceiling effect: NS

SF-36 SF *Eng*[28]	Poor/?							Poor/?	
Results: Box A: Unidimensionality not checked/reported. Cronbach α (0.80-0.91), no exact value was reported. Box I: Responsiveness is tested in different ways but no evidence for responsiveness was achieved (small sample size of subanalysis and or insufficient methods applied). **Interpretability**: N=70, Missing data: 12.5% (excl.). Mean(SD) score at baseline/follow-up: 81.57(24.09)/ 67.27(25.79)Floor/ceiling: NS.									
SF-36 SF *Chin*[40]	Poor/?					Good/+			F/C
Results: Box A: Unidimensionality not sufficiently checked. Cronbach α 0.787. Box F: Convergent validity (internal relationships): Scaling assumption (equal item variance, item-own scale, item-other scale) in consistency with hypotheses but for known group validity (external relationships) the hypotheses were vaguely stated. **Interpretability:** N=168. Missing data: NS. Mean(SD) scale score: 66.42(26.28) Floor effect: 1.2%, Ceiling effect: 17.6%.									
SF-36 SF *Chin*[50]								Poor/?	MID
Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in SF-36 Soc.F score: r_s= -0.27. **Interpretability:** N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Patients treated > 12 weeks: 50.1(29.2) and 2) Patients treated <12 weeks: 55.9(31.3). MID for improvement: 5.06(22.50). MID for deterioration: -2.78(17.45). Effect size: 0.26, SRM: 0.31. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS									
EMOTIONAL WELL BEING	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
SF-36 MH *Eng*[26]	Poor/?					Fair/?			
Results: Box A: Limited information on unidimensionality. Cronbachα 0.87. Box F: Sparse information on properties of comparators, convergent validity assessed by comparing to unrelated clinical measures (function, disease activity, disease severity). Known group validity assessed by comparison of scores from general population and PsA, showing no statistically significant difference. However, this is not enough information to generate a negative score for box F. **Interpretability:** N=113. Missing data: NS (all completed) Mean(SD) scale score: 73.0(2.21). Floor/Ceiling effect: NS									
SF-36 MH *Eng*[28]	Poor/?							Poor/?	
Results: BOX A: Unidimensionality not checked/reported. Cronbach α (0.80-0.91), no exact value was reported. Box I: Responsiveness is tested in different ways but no evidence for responsiveness was achieved (small sample size of subanalysis and or insufficient methods applied). **Interpretability**: N=70, Missing data: 12.5% (excl.). Mean(SD) score at baseline/follow-up: 73.6(18.64)/ 67.66(23.90). Floor/ceiling: NS.									
SF-36 MH *Chin*[40]	Poor/?					Good/+			
Results: Box A: Unidimensionality not sufficiently checked. Cronbach α = 0.808. Box F: Convergent validity (internal relationships): Scaling assumption (equal item variance, item-own scale, item-other scale) in consistency with hypotheses but for known group validity (external relationships) the hypotheses were vaguely stated. **Interpretability:** N=168. Missing data: NS. Mean(SD) scale score: 63.95(19.65) Floor effect: 0%, Ceiling effect:3 %.									
SF-36 MH *Chin*[50]								Poor/?	MID
Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in SF-36 MH score: r_s= -0.28. **Interpretability:** N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Patients treated > 12 weeks: 50.7 (23.1) and 2) Patients treated									

<12 weeks: 57.5(11.5). MID for improvement: 1.41(8.36). MID for deterioration: -0.00(14.85). Effect size: 0.19, SRM: 0.31. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS			
SF-36 MCS *Chin*[40]	Good/+	Poor/?	
Results: Box E: Structural validity assessed by PCA, and a 2 factor model (PCS, MCS) was supported, explaining 69.4% of the total variance. Box F: Only known group validity (general population vs. PsA) and no exact hypotheses stated a priori. **Interpretability:** N=168. Missing data: NS. Mean(SD) component summary score: 45.22(12.66.)			
SF-36 MCS *Chin*[50]		Poor/?	MID
Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in SF-36 mcs: r_s= -0.24 **Interpretability:** N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Patients treated > 12 weeks: 38.8 (12.6) and 2) Patients treated <12 weeks: 38.8(9.2). MID for improvement: 1.77(8.60). MID for deterioration: -0.70(8.37). Effect size: 0.19, SRM: 0.28. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS			
AIMS1 Psyc.C. *Eng*[25]		Poor/?	
Box I: Correlation between change in clinical measures and scores of AIMS1 (time 1) and AIMS2 (time 2). Not assessing correlation between change scores and no a priori hypotheses. **Interpretability:** N = 65. Missing data: NS. Mean(SD) for scale score: 3.34(1.04). Time frame (responsiveness): 4 years. Floor/ceiling effect: NS			
AIMS1 Anxiety *Eng*[21]		Fair/?	
Result: Box F: Sparse information about properties of comparators and all comparators represent unrelated constructs (function, disease activity or severity) why no results score is generated. **Interpretability:** N=145. Missing data: NS Mean(SD) scale score: 1.7(0.7). Floor/ceiling effect: NS			
AIMS1 Depression *Eng*[21]		Fair/?	
Result: Box F: Sparse information about properties of comparators and all comparators represent unrelated constructs (function, disease activity or severity) why no result is generated. **Interpretability:** N=145. Missing data: NS. Mean(SD) scale score: 2.4(1.0). Floor/ceiling effect: NS			
AIMS2 Mood *Eng*[24]		Fair/?	
Result: Box F: Sparse information about properties of comparators and all comparators represent unrelated constructs (function, disease activity or severity) why no result is generated. **Interpretability:** N=124. Missing data: NS Mean(SD) scale score: 2.67(1.60). Floor/ceiling effect: NS			
AIMS2 Tension *Eng*[24]		Fair/?	
Result: Box F: Sparse information about properties of comparators, and all comparators represent unrelated constructs (function, disease activity or severity) why no result is generated. **Interpretability:** N=124. Missing data: NS Mean(SD) scale score: 4.32(1.93). Floor/ceiling effect: NS**:**			
AIMS2 Psyc.C. *Eng*[28]		Poor/?	
Results: Box I: Responsiveness is tested in different ways but no evidence for responsiveness was achieved (small sample size of subanalysis and or			

insufficient methods applied). **Interpretability:** N = 80. Missing data: 12% (excl.). Mean(SD) for scale scores at 1st/2nd visit: 3.68(1.67)/3.69(2.00). Time interval 12-18 months. Floor/ceiling effect: NS

AIMS2 Psyc.C. *Eng*[25] Poor/?

Result: Box I: Correlation between change in clinical measures and scores of AIMS1 (time 1) and AIMS2 (time 2). Not assessing correlation between change scores and no a priori hypotheses.
Interpretability: Mean(SD) scale score: 3.98(2.69) (2nd assessment, 1st assessment was AIMS1, 4 years interval.)

mRAI (of MultiP) *Eng*[44] Poor/? Poor/? Poor/?

Results: Box B: Test-retest: Number and % PsA patients in analysis was not specified. ICC (95%CI) = 0.94 (0.93-0.94) Box D: Not enough information available to rate quality or result of content validity assessment. Box F: Sparse hypotheses and no information about measurement properties of comparators and for most correlations, the proportion of PsA was <50%. Correlation between mRAI to DAS28 (r = 0.741) (%PsA not clear) and between mRAI and patient reported TJC: r=0.672 (PsA n=57).
Interpretability: Total N = 462 (PsA 123, 26.6%). Number of PsA patients not reported for all analyses. Missing data: NS. Mean(SD) baseline score: 6(0.47) (26.6% PsA). Time frame test-retest: 1 week. Floor/ceiling effect: NS.

ECONOMIC COST

EQ-5D-3L *Norw*[45] MCII, PASS

Interpretability: Total number of PsA in the study: N = 1391. No. of PsA in the analysis: n = 250. Missing data: Reported per analysis (excl.). Mean(SD) utility score: 0.49(0.29) at baseline and 0.61(0.28) after 3 months. The anchoring questions were given after 3 months and asked about satisfaction (PASS) and improvement (MCII), respectively. PASS cut-points (ROC curves): 75% sensitivity cut-off: 0.69. 80% specificity cut-off: 0.73. AUC (95%CI): 0.78 (0.72-0.84). MCII cut-points (ROC curves): 75% sensitivity cut-off: 0. 80% specificity cut-off: 0.18. AUC(95%CI): 0.678(0.61-0.75).

EQ-5D-3L *Swe[75]* PASS

Interpretability: N=255 (PsA=23). Objective of study was to compare British (UK), hypothetical, and Swedish (SE), experience-based, EQ-5D utilities using data from clinical practice/cohort of patients with RA, SpA and PsA treated with anti-TNFI. Point estimates and PASS cut-off levels were compared: SE utilities were higher than UK utilities: Baseline mean(SD) UK/SE EQ-5D: 0.44(0.34)/0.72(0.15). PASS cut offs were stable over time for both the UK and SE preference: Baseline: Mean(SD) UK/SE EQ-5D: 0.71/0.84, follow-up: 0.71/0.82 but higher (0.11) when using SE compared to UK. Percentage in PASS at baseline/follow-up (18.9 % /59.9 %).

EQ-5D-3L *Hung*[42] Fair/+

Results: Box F: Sparse hypotheses and information about measurement properties of comparators. Moderate/strong correlation (r = 0.63-0.73) with related measures (PsAQoL, HAQ, BASDAI, PGA, VAS pain). Known group validity: Differences in scores according disease severity, SMD: 0.46 to 1.1.
Interpretability. N = 183. Missing data: 3%. Mean(SD) scores for EQ-5D utility: 0.5 (0.3). Median(range) score: 0.587 (–0.594 to 1). Scores for EQ-5D VAS: Mean(SD) score: 54.7 (20.0), median(range) score: 52(5–95). Floor/ceiling effect: NS.

EQ-5D-3L *Eng/Chin*[54] Fair/+ F/C

Results: Box F: Vague hypotheses. Moderate correlation between EQ-5D utility score and related measures: SF-6D utility (r_s= 0.594), SF-GH (r_s = -0.44), PCS (r_s= 0.445), MCS (r_s= 0.371), EQ-VAS (r_s= 0.494). Known group validity: Differences in scores according to SF-general health status, Effect size

ranging from 0.62 (poor/fair vs good health state) to 0.91 (excellent/very good vs good health state).
Interpretability: N = 86. Missing data: 1.2%. Mean(SD) score: 0.74(0.24), median(IRQ): 0.8(0.09). Bimodal score distribution. Floor effect: 2.3% had negative scores for EQ-5D. Ceiling effect: 20%.

SF-6D *Eng/Chin*[54] Fair/+ F/C

Results: Box F: Vague hypotheses. Moderate/strong correlation to related measures: EQ-5D utility (r_s= 0.594), SF-GH (r_s=-0.569), PCS (r_s= 0.843), MCS (r_s=0.623), EQ-VAS (r_s= 0.538). Known group validity: Differences in scores according to SF-general health status, ES ranging from 0.92 (poor/fair vs good health state) to 0.94 (excellent/very good vs good health state).
Interpretability: N = 86. Missing data: 9.3% for SF-6D, reduced to 3.5% by estimating the missing by SF-36v2 protocol. Mean(SD) score: 0.68(0.13), median (IRQ): 0.64(0.18). Normal score distribution. Floor/ceiling effect: None.

SF-6D *Norw*[45]

Interpretability: Total number of PsA in the study: N = 1391. No. of PsA in the analysis: n = 819. Missing data: Stated for each part of the study. Mean(SD) scores utility: 0.60(0.12) at baseline and 0.66(0.13) at 3 months. Anchoring questions about satisfaction (PASS) and improvement (MCII), respectively. PASS cut-points (ROC curves): 75% sensitivity cut-off: 0.60. 80% specificity cut-off: 0.65. AUC 0.80(95%CI 0.76-0.82). MCII cut-points (ROC curves): 75% sensitivity cut-off: 0.01. 80% specificity cut-off: 0.07. AUC(95%CI): 0.73(0.69-0.76). Floor/ceiling effect: NS

EQ-5D and SF-6D *Eng[47]* Poor/? Poor/? Score distributions

Results/Interpretability: Box F, I: The study compares utility estimates obtained from EQ-5D and SF-6D mapped to HAQ. No a priori hypotheses are stated about expected correlations, and no gold standard explained. It is only possible to conclude that the measurements are not similar no conclusion about the measurement properties for each of them can be drawn. Change in utility score during 1 year of biologic treatment is 0.09 for SF-6D and 0.28 for EQ-5D. EQ-5D-derived utilities are likely to produce larger QALY gains than SF-6D-derived utilities for a given change in the disease-specific measure (HAQDI). The EQ-5D displays a bimodal distribution in more severe health states in both RA and PsA. Mean(utility scores): SF-6D: 0.57(0.12) at baseline and 0.66(0.12) at follow up (1 year). EQ-5D: 0.49(32) at baseline and 0.77(28) at follow up. Utility scores were calculated from the preference-based instruments using UK population norms.

EQ-5D-rev. *Eng**[48]*** Poor/? Poor/? Score distributions

Results/interpretability: A revised scoring of EQ-5D is used and shown to lessen the gap between utility measures produced by SF_6D and the original EQ-5D utility estimates, regression analysis with HAQ-DI as independent variable show more comparable slope with the revised EQ-5D and the SF-6D compared to the original. However it is not possible to provide a score for the construct validity of the EQ_5D revised version because no hypotheses are stated a priori about expected correlations. Baseline mean(SD)/range scores original vs revised: 0.49(32)/-0.24;1.0 vs 0.62(0.21)/-0.14;0.99. Followup: 0.77(28)/-0.24;1.0) vs 0.84(0.17)/0.046;0.9954). Change in utility 0-12 months (biologic treatment) 0.28(-0.36;0.2) vs 0.22 (0.28;0.167)

WTP *Eng*[46] Poor/? Fair/+

Results: (Pilot study). Box D: Content validity of the tool was reviewed by rheumatologists not PsA patients during the development phase. The

proportion of patients confirming an impact of PsA on these domains varied between 35%-88%. Box F: Correlation between median WTP amounts were higher between related than non-related domains. However, one of the 4 domains that patients ranked as most impacted by PsA was associated with a lower WTP amount than some domains ranked at less impacted by PsA.
Interpretability: N = 60. Missing data: NS. Median(IQR) scores of WTP for relief of the 8 domains: Lowest WPT amount (concentration): 7500(1000-50,000), highest WTP amount (Physical comfort, Sleep, work): 10,000(5000-75,000).

SLEEP	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
VAS sleep *Eng*[43]										MID

Interpretability: N = 200. Missing data: NS. Mean(SD) (1st/2nd visit): 37.99(32.93)/38.83(32.32). Varying time interval (mean: 8.28 months between visits). MID(SD) estimates for improvement/worsening: -10.97(29.74)/13.96(27.32). Correlation between mean change in VAS pain and anchor (patient's rating of change): $r_s = 0.326$ (the authors did not aim to test responsiveness of VAS pain, only MID) Floor/ceiling effect: NS.

STIFFNESS	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
NRS stiffness *Eng*[44]				Poor/?		Poor/?			

Results: Box D: Not enough information to rate quality or results on content validity. Box F: No hypotheses or information about properties of comparators, PsA <50% of the population in most of the analyses. Correlation between patient-reported TJC and NRS stiffness (r=0.600) (in PsA subset n=57).
Interpretability: Total N = 462 (PsA 123, 26.6%). Number of PsA patients not reported for all analyses. Missing data: NS. Mean(SD)score (baseline): NS. Floor/Ceiling effect: NS.

VAS stiffness *Eng*[22]						Poor/?				

Results: Box F: No hypotheses and sparse information about measurement properties of comparators. Moderate correlation to ACR functional class not to other measures presented.
Interpretability: N = 99-114. Missing data: 13% (excl.). Mean(SD) score: 0.91(0.69). Floor/ceiling effect: NS

NON-COS Domains	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
SF-36 GH *Eng*[26]	Poor/?					Fair/-				

Results: Box A**:** Cronbachα 0.82. Limited evidence for unidimensionality: Less than 75% of convergent validity hypotheses confirmed about correlation to measures of function, disease activity and severity. Known group validity showing significant difference to general population but not enough information to generate a positive score for box F.
Interpretability: N=113. Missing data: NS (all completed) Mean(SD) scale score: 58.8(3.31). Floor/Ceiling effect: NS

SF-36 GH *Chin*[40]	Poor/?					Good/-				F/C

Results: Box A: Unidimensionality not reported, Cronbach α = 0.749. Box F: Hypotheses about internal relationships (scaling assumptions) not sufficiently fulfilled item 1 of GH had higher other-scale (BP, RP, PF, VT scales) than own-scale correlations. Known group validity assessed (according to HAQ, BASDAI, DAS28 level) with higher SF-36GH scores in the severe groups but no hypotheses (about expected magnitude of difference) stated a priori.
Interpretability: N=168. Missing data: NS (all completed). Mean(SD) score 41.53(21.00). Floor effect: 2.4. Ceiling effect: 0.

SF-36 GH *Chin*[50]	Poor/?	MID
Results: Box I: Small sample size. Correlation between anchor (patient perception of change) and change in SF-36 GH: r_s= -0.30 **Interpretability:** N = 17-21 in analyses. Missing data: NS. Mean(SD) baseline scores: 1: Patients treated > 12 weeks: 23.8 (8.0) and 2) Patients treated <12 weeks: 40.5(11.0). MID for improvement: -2.94(12.34). MID for deterioration: -0.3.75(15.02). Effect size: 0.25, SRM: 0.24. Time frame (responsiveness): up to 52 weeks. Floor/ceiling effect: NS		
AIMS-2 Social Support *Eng*[24]	Fair/?	
Results: Box F: Sparse information about properties of comparators and only correlations to unrelated measures (disease activity, severity and function) Interpretability: N=124. Missing data: NS. Mean(SD) score: 1.82(1.86). Floor/Ceiling effect: NS		

ACR20/ACR50/ACR70: American College of Rheumatology response criteria (20/50/70% improvement); ARA; American Rheumatism Association; ASDAS, Ankylosing Spondylitis Disease Activity Score;AUC, Area Under Curve; BSA; Body Surface Area (with psoriasis); CDAI, Clinical Disease Activity Index; cDAPSA, clinical Disease Activity index for Psoriatic Arthritis; Chin, Chinese; CPDAI, Composite Psoriatic Disease Activity Index; DAS28; Disease Activity Score-28 joints; DLQI, Dermatology Life Quality Index; Eng, English; ERS; Erythrocyte sedimentation rate; Excl, Excluded; FA, Factor Analysis; F/C, Floor and/or Ceiling effect reported; FI, Functional Index; Germ, German; Hung, Hungarian; DIF, Differential Item Functioning; ICC, Inter-correlation coefficient; IPBOD, Inverse Psoriasis Burden of Disease questionnaire; Ital, Italian; IQR, Interquartile range; LoA, Limits of agreement; MCID, Minimal Clinically Important Difference; MCII, Minimal clinical important improvement; MDA, Minimal Disease Activity; MDC, minimal detectable change; MIC, Minimal important change; MID, Minimal important Difference; Missing data (either item responses or patients); N, Number of patients; NHP(D); Nottingham Health Profile(Distress index); Norw, Norwegian; NS, Not Stated; PASDAS, Psoriatic Arthritis Disease Activity Score; PASI; Psoriasis Activity and Severity Index; PASS, Patient acceptable symptom state; PCA, Principal component analysis; PGA, Patient Global Assessment; PhGA, Physician Global Assessment; PSD, Psoriasis Symptom Diary; PSI, Psoriasis Symptom Inventory; r_s, Spearman correlation coefficient; r; Pearson correlation coefficient; RA, Rheumatoid arthritis; ROC, Receiver Operating Curve; SD, Standard deviation; SJC, Swollen Joint Count; SMD, Standard Mean Difference; SpA; spondyloarthropathy; Span, Spanish; SRM, Standard Response Mean; Swe, Swedish; TJC, Tender Joint Count; Turk, Turkish; VAS, Visual Analogue Scale.

Supplementary Table E: Best-evidence synthesis of measurement properties for each instrument *evaluated separately for each language version*

PROMs/scales listed according to COS Domain category	Reliability COSMIN BOX (A-C)			Validity COSMIN BOX (D-H)					Responsiveness COSMIN BOX (I)	Relevant info on score interpretation reported?
	Internal consistency	Relia-bility	Measure-ment error	Content validity	Structural validity	Hypothe-ses testing	Cross-cult. Validity	Criterion validity	Sensitivity to change	
	A	B	C	D	E	F	G	H	I	
MSK DISEASE ACTIVITY										
BASDAI *Eng*						±				F/C
BASDAI *Span*	?					–			?	F/C
SASPA *Germ*	+				+	?			?	
PASE-total *Eng*		?				?			?	
PASE-symptom *Eng*		?				?			?	
PASE-function *Eng*		?				?			?	
PASE-total *Ital*						+	a		+	
PASE-symptom *Ital*						+	a		+	
PASE-function *Ital*						+	a		+	
PR-TJC (MultiP) *Eng*				?		?				
SKIN DISEASE ACTIVITY										
PSI *Eng*	++	+			++	+			+	F/C
PSD *Eng*				+++						
Worst itch NRS *Eng*				+						
PAIN										
VAS pain (recall 1 week) *Eng*						+			?	MID
VAS pain (recall NS)										PASS, MCII

Norw							
VAS pain *Chin*						?	MID
NRS pain *Eng*		?	?	?			
SF-36 BP *Eng*	?			+		?	
SF-36 BP Chin	?			++ b		?	MID, F/C
AIMS1 Pain *Eng*				+		?	
AIMS1 Pain *Ital*				+			
AIMS2 Pain *Eng*	?			+		?	
PATIENT GLOBAL							
Patient Global due to psoriasis							
NRS skin impact (1 week recall) *Eng*				+			F/C
VAS skin impact (1 week recall) *Eng*		++		?			
Patient global due to arthritis							
NRS joint impact (1 week recall)*Eng*			?	+			F/C
NRS joint impact (1 day recall) *Eng*							
VAS joint impact *Eng*		++		?			
Patient global due to PsA							
PGA by NRS (1 week recall) *Eng*				+			
PGA by NRS (1 week recall) *Chin*				+			F/C
PGA by VAS (1 week recall) *Eng*		++		?			MID
PGA by VAS (1 week recall) *Ital*				?	?		
PGA by VAS (unknown recall) *Norw*							PASS, MCII
PGA by VAS (1 week recall) *Chin*						?	MID
PHYSICAL FUNCTION							
DFI *Chin*	– –		– –	?			
DASH *Eng*				– –			
BASFI *Chin*	++		++	?			F/C
HAQ-DI *Eng*	++		++	–		?	F/C, MID
HAQ-DI *Ital*				–			

HAQ-DI *Chin*	++				++	?			?	F/C
HAQ-DI *Hung*						+				
HAQ-DI *Thai*	?					+				F/C
HAQ-S *Eng*						–				
HAQ-SK *Eng*						?				
mHAQ *Norw*										PASS, MCII
SF-36 PF *Eng*	++				++	+			?	F/C
SF-36 PF *Chin*	++				++	++ b			?	F/C
SF-36 PCS *Chin*					++	?			?	
CASQ-FI *Eng*	?				?	?				
AIMS1 Mobility *Eng*						-				
AIMS1 Physical *Eng*						+				
AIMS1 Dexterity *Eng*						+				
AIMS1 House *Eng*						+				
AIMS1 Physical *Ital*						-				
AIMS1 ADL *Eng*						-				
AIMS1 PC. *Eng*									?	
AIMS2 PC. *Eng*									?	
AIMS2 Mobility *Eng*						+				
AIMS2 Physical *Eng*						+				
AIMS2 Dexterity *Eng*						+				
AIMS2 Selfcare *Eng*						-				
AIMS2 House. *Eng*						-				
AIMS2 Arm F. *Eng*						+				
HRQoL/MULTIDIM. PROMs	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**Interpretability**
PsAQoL *Eng*	++	?		+++	++	+			?	
PsAQol *Eng/Chin*	?	+		?		+	a			
PsAQoL *Swe*	++	?		++		+	a			F/C
PsAQoL *Hung*						+				

PsAQoL *Dutch*	?	++	?	−−		+	a			
AIMS1 Global *Ital*						?				
PsAID-9 *Eng*	c	++		+++		+	*a*		?	PASS, F/C
PsAID-12 *Eng*	c	++		+++		+	*a*		?	PASS, F/C
PsAID-12touch *Ital*						+		+		MDA cut-off
PsAID-12 *Ital*	c				c	+				Cut-off values
PAIP *Ital*						?				
VITACORA-19 *Span*	+	++		++	+	+			?	MCID, F/C
VITACORA-19 *Ital*	?	+			?	+	a			
PsoDisk *Ital*									?	
CIAQ-QoL *Eng*		?		?		?				
IPBOD *Eng*	?			?		?				
FATIGUE										
FACIT-Fatigue *Eng*	?	+				+				
NRS fatigue *Eng*		?		?		?			?	
VAS fatigue *Eng*										MID
SF-36 VT *Eng*	?					-				
SF-36 VT *Chin*	?					++b			?	MID, F/C
PARTICIPATION										
SRPQ IM. *Eng*	?	+	?	+		+				MDC
SRPQ ST *Eng*	?	+	?	+		+				MDC
SRPQ SR *Eng*	?	+	?	+		+				MDC
WPS *Eng*						+			+	F/C
AIMS1 SA *Eng*						?				
AIMS2 SA *Eng*						?				
AIMS2 Work *Eng*						?				
AIMS2 SC *Eng*									?	
SF-36 RE *Eng*	?					?				
SF-36 RE *Chin*	?					++ b			?	
SF-36 RP *Eng*						-				

SF-36 RP *Chin*		?			++ b	?	
SF-36 SF *Eng*		?			?	?	
SF-36 SF *Chin*		?			++ b	?	
EMOTIONAL WELL-BEING							
SF-36 MH *Eng*	?				?	?	
SF-36 MH *Chin*	?				++ b	?	MID
SF-36 MCS *Chin*						?	
SF-36 MCS *Chin*				++	?	?	
mRAI *Eng*		?	?		?		
AIMS1 Psyc.C. *Eng*						?	
AIMS1 Anxiety *Eng*					?		
AIMS1 Depression *Eng*					?		
AIMS2 Mood *Eng*					?		
AIMS2 Tension *Eng*					?		
AIMS2 Psyc.C. *Eng*						?	
ECONOMIC COST							
EQ-5D-3L *Norw*							MCII, PASS
EQ-5D-3L *Swe*							PASS
EQ-5D-3L *Hung*					+		
EQ-5D-3L*Eng/Chin*				+			F/C
EQ-5D-3L *Eng*					?	?	Utility score distribution
EQ-5D-3L-rev. *Eng*							
SF-6D *Eng*					?	?	Utility score distribution
SF-6D *Eng/Chin*					+		F/C
SF-6D Norw							PASS, MCII,
WTP *Eng*			?		+		
SLEEP							
VAS sleep *Eng*							MID
STIFFNESS							

NRS stiffness *Eng*		?	?			
HAQ VAS Stiffness *Eng*			?			
NON-COS Domains						
SF-36 GH *Eng*	?		–			
SF-36 GH *Chin*	?		– –		?	MID,F/C
AIMS2 Social Support *Eng*			?			

Empty cells reflect that the measurement properties were not evaluated by any study for the given instrument**. T**able 2 explains the grading of evidence (+/-/?).[a]Only translation, no cross-cultural validation. According to COSMIN, only studies that address measurement invariance (e.g. multiple group factor analyses or DIF) between countries (or other groups) are considered real cross-cultural validity studies [b]Construct validity – hypotheses testing was assessed regarding the internal relationships (scale assumptions) and not external measures.[c]Questionnaire based on a formative model why internal consistency and structural validity are not rated. AIMS, Arthritis Impact Measurement Scales (ADL, Activity of daily living; Arm F., Arm Function; House, Household; PC, Physical component score; Psyc.C., Psychological component score; SA, Social Activity, SC, Social component score); BASDAI, Bath Ankylosing Spondylitis Activity Index; BASFI, Bath Ankylosing Spondylitis Functional index; Chin, Chinese; CIAQ-FI, Combined Inflammatory Arthritis – Functional Impairment questionnaire; CIAQ-Qol, Combined Inflammatory Arthritis – quality of life questionnaire; COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; DASH, Disabilities of the Arm Shoulder and Hand Outcome Measure; DFI, Dougados Functional Index; Eng, English; EQ-5D-3L, EuroQoL 5 Dimensions questionnaire with 3 response levels; FACIT-Fatigue, Functional Assessment of Chronic Illness Therapy-Fatigue scale; F/C, Floor/Ceiling effect; Germ, German; HAQ, Health Assessment Questionnaire (HAQ-S: Spondyloarthropathy, HAQ-SK: Skin, HAQ-DI: Disability Index); Hung, Hungarian; IPBOD, Inverse Psoriasis Burden of Disease questionnaire; Ital, Italian; MCID, minimal clinically important difference; MDC, minimal detectable change; MCII, minimal clinical important improvement; MDA, minimal disease activity; MIC, minimal important change; MID, minimal important difference; mRAI, Modified Rheumatology Attitude Index; MultiP, Multidimensional Patient Reported Outcome Questionnaire; Norw, Norwegian; NRS, Numeric Rating Scale; PAIP, Psoriatic Arthritis Impact Profile; PASE, PsA Screening and Evaluation Questionnaire; PASS, Patient acceptable symptom state; PGA, Patient Global Assessment; PR-TJC, Patient-reported-tender-joint-count; PsAID, Psoriatic Arthritis Impact of Disease questionnaire; PsAQoL, PsA Quality of Life instrument; PSD; Psoriasis Symptom Diary; PSI, Psoriasis Symptom Inventory; PsoDisk questionnaire, no full spelling available; SASPA, Stockerau Activity Score for Psoriatic Arthritis; SF-6D, utility tool derived from SF-36 comprising six multi-level dimensions; SF-36, Medical Outcome Survey Short Form 36-item Health Survey (SF-36 subscales: BP, Bodily Pain; GH, General Health; MCS, Mental Component Summary; MH, Mental Health; PCS, Physical Component Summary, PF, physical function; RE, Role Emotional; RP, Role Physical; SF, Social Functioning; VT, Vitality); Span, Spanish; SRPQ, Social Role Participation Questionnaire; Swe, Swedish; VAS, Visual Analogue Scale; VITACORA-19, Spanish acronym, full name not available; WTP, Willingness to Pay questionnaire; WPS, Work Productivity Survey.

Reference List

(1) Gladman DD, Antoni C, Mease P *et al.* Psoriatic arthritis: epidemiology, clinical features, course, and outcome. *Ann Rheum Dis* 2005;**64 Suppl 2**:ii14-ii17.

(2) Gulati AM, Semb AG, Rollefstad S *et al.* On the HUNT for cardiovascular risk factors and disease in patients with psoriatic arthritis: population-based data from the Nord-Trondelag Health Study. *Ann Rheum Dis* 2015.

(3) Ogdie A, Schwartzman S, Husni ME. Recognizing and managing comorbidities in psoriatic arthritis. *Curr Opin Rheumatol* 2015;**27**:118-26.

(4) Gladman DD, Mease PJ, Strand V *et al.* Consensus on a core set of domains for psoriatic arthritis. *J Rheumatol* 2007;**34**:1167-70.

(5) Orbai AM, de WM, Mease P *et al.* International patient and physician consensus on a psoriatic arthritis core outcome set for clinical trials. *Ann Rheum Dis* 2016.

(6) Boers M, Kirwan JR, Wells G *et al.* Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. *J Clin Epidemiol* 2014;**67**:745-53.

(7) Kalyoncu U, Ogdie A, Campbell W *et al.* Systematic literature review of domains assessed in psoriatic arthritis to inform the update of the psoriatic arthritis core domain set. *RMD Open* 2016;**2**:e000217.

(8) Mokkink LB, Terwee CB, Patrick DL *et al.* The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;**63**:737-45.

(9) Martin Boers, John Richard Kirwan, Peter Tugwell *et al.* The OMERACT handbook. 1-5-0016.
Ref Type: Online Source

(10) Boers M, Brooks P, Strand CV *et al.* The OMERACT filter for Outcome Measures in Rheumatology. *J Rheumatol* 1998;**25**:198-9.

(11) Terwee CB, Mokkink LB, Knol DL *et al.* Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012;**21**:651-7.

(12) De Vet H.C.W, Terwee C.B., Mokkink L.B, Knol D.L. *Measurement in Medicine*. 2 ed. Cambridge University Press, 2013.

(13) Cosmin.nl. **CO**nsensus-based **S**tandards for the selection of health **M**easurement **IN**struments. 26-9-2015.
Ref Type: Online Source

(14) Liberati A, Altman DG, Tetzlaff J *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med* 2009;**151**:W65-W94.

(15) Terwee CB, Jansma EP, Riphagen II *et al.* Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res* 2009;**18**:1115-23.

(16) Castrejón I GLCL, et al. EULAR outcome measures library. 2016.

Ref Type: Online Source

(17) Prinsen CA, Vohra S, Rose MR *et al.* How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. *Trials* 2016;**17**:449.

(18) van TM, Furlan A, Bombardier C *et al.* Updated method guidelines for systematic reviews in the cochrane collaboration back review group. *Spine (Phila Pa 1976)* 2003;**28**:1290-9.

(19) Gerbens LA, Prinsen CA, Chalmers JR *et al.* Evaluation of the measurement properties of symptom measurement instruments for atopic eczema: a systematic review. *Allergy* 2017;**72**:146-63.

(20) US Department of Health and Human Services. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. (2009). 2009.

Ref Type: Online Source

(21) Duffy CM, Watanabe Duffy KN, Gladman DD *et al.* The utility of the arthritis impact measurement scales for patients with psoriatic arthritis. *J Rheumatol* 1992;**19**:1727-32.

(22) Blackmore MG, Gladman DD, Husted J *et al.* Measuring health status in psoriatic arthritis: the Health Assessment Questionnaire and its modification. *J Rheumatol* 1995;**22**:886-93.

(23) Husted JA, Gladman DD, Long JA *et al.* A modified version of the Health Assessment Questionnaire (HAQ) for psoriatic arthritis. *Clin Exp Rheumatol* 1995;**13**:439-43.

(24) Husted J, Gladman DD, Farewell VT *et al.* Validation of the revised and expanded version of the Arthritis Impact Measurement Scales for patients with psoriatic Arthritis. *J Rheumatol* 1996;**23**:1015-9.

(25) Husted J, Gladman DD, Long JA *et al.* Relationship of the Arthritis Impact Measurement Scales to changes in articular status and functional performance in patients with psoriatic arthritis. *J Rheumatol* 1996;**23**:1932-7.

(26) Husted JA, Gladman DD, Farewell VT *et al.* Validating the SF-36 health survey questionnaire in patients with psoriatic arthritis. *J Rheumatol* 1997;**24**:511-7.

(27) Taccari E, Spadaro A, Rinaldi T *et al.* Comparison of the Health Assessment Questionnaire and Arthritis Impact Measurement Scale in patients with psoriatic arthritis. *Rev Rhum Engl Ed* 1998;**65**:751-8.

(28) Husted JA, Gladman DD, Cook RJ *et al.* Responsiveness of health status instruments to changes in articular status and perceived health in patients with psoriatic arthritis. *J Rheumatol* 1998;**25**:2146-55.

(29) Navsarikar A, Gladman DD, Husted JA *et al.* Validity assessment of the disabilities of arm, shoulder, and hand questionnaire (DASH) for patients with psoriatic arthritis. *J Rheumatol* 1999;**26**:2191-4.

(30) McKenna SP, Doward LC, Whalley D *et al.* Development of the PsAQoL: a quality of life instrument specific to psoriatic arthritis. *Ann Rheum Dis* 2004;**63**:162-9.

(31) Taylor WJ, Harrison AA. Could the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) be a valid measure of disease activity in patients with psoriatic arthritis? *Arthritis Care and Research* 2004;**51**:311-5.

(32) Chandran V, Bhella S, Schentag C *et al.* Functional assessment of chronic illness therapy-fatigue scale is valid in patients with psoriatic arthritis. *Ann Rheum Dis* 2007;**66**:936-9.

(33) Taylor WJ, McPherson KM. Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. *Arthritis Rheum* 2007;**57**:723-9.

(34) Leung Y-Y, Tam L-S, Kun EWL *et al.* Comparison of 4 functional indexes in psoriatic arthritis with axial or peripheral disease subgroups using Rasch analyses. *Journal of Rheumatology* 2008;**35**:1613-21.

(35) Healy PJ, Helliwell PS. Psoriatic arthritis quality of life instrument: an assessment of sensitivity and response to change. *J Rheumatol* 2008;**35**:1359-61.

(36) Dominguez PL, Husni ME, Holt EW *et al.* Validity, reliability, and sensitivity-to-change properties of the psoriatic arthritis screening and evaluation questionnaire. *Arch Dermatol Res* 2009;**301**:573-9.

(37) Fernandez-Sueiro JL, Willisch A, Pertega-Diaz S *et al.* Validity of the bath ankylosing spondylitis disease activity index for the evaluation of disease activity in axial psoriatic arthritis. *Arthritis Care Res (Hoboken)* 2010;**62**:78-85.

(38) Minnock P, Kirwan J, Veale D *et al.* Fatigue is an independent outcome measure and is sensitive to change in patients with psoriatic arthritis. *Clin Exp Rheumatol* 2010;**28**:401-4.

(39) Eder L, Chandran V, Shen H *et al.* Is ASDAS better than BASDAI as a measure of disease activity in axial psoriatic arthritis? *Ann Rheum Dis* 2010;**69**:2160-4.

(40) Leung YY, Ho KW, Zhu TY *et al.* Testing scaling assumptions, reliability and validity of medical outcomes study short-form 36 health survey in psoriatic arthritis. *Rheumatology (Oxford)* 2010;**49**:1495-501.

(41) Billing E, McKenna SP, Staun M *et al.* Adaptation of the Psoriatic Arthritis Quality of Life (PsAQoL) instrument for Sweden. *Scand J Rheumatol* 2010;**39**:223-8.

(42) Brodszky V, Pentek M, Balint PV *et al.* Comparison of the Psoriatic Arthritis Quality of Life (PsAQoL) questionnaire, the functional status (HAQ) and utility (EQ-5D) measures in psoriatic arthritis: results from a cross-sectional survey. *Scand J Rheumatol* 2010;**39**:303-9.

(43) Kwok T, Pope JE. Minimally important difference for patient-reported outcomes in psoriatic arthritis: Health Assessment Questionnaire and pain, fatigue, and global visual analog scales. *J Rheumatol* 2010;**37**:1024-8.

(44) El MY, El GM, Youssef SS *et al.* Incorporating patient reported outcome measures in clinical practice: Development and validation of a questionnaire for inflammatory arthritis. *Clinical and Experimental Rheumatology* 2010;**28**:734-44.

(45) Kvamme MK, Kristiansen IS, Lie E *et al.* Identification of cutpoints for acceptable health status and important improvement in patient-reported outcomes, in rheumatoid arthritis, psoriatic arthritis, and ankylosing spondylitis. *J Rheumatol* 2010;**37**:26-31.

(46) Hu SW, Holt EW, Husni ME *et al.* Willingness-to-pay stated preferences for 8 health-related quality-of-life domains in psoriatic arthritis: a pilot study. *Semin Arthritis Rheum* 2010;**39**:384-97.

(47) Adams R, Walsh C, Veale D *et al.* Understanding the relationship between the EQ-5D, SF-6D, HAQ and disease activity in inflammatory arthritis. *PharmacoEconomics* 2010;**28**:477-87.

(48) Adams R, Craig BM, Walsh CD *et al.* The impact of a revised EQ-5D population scoring on preference-based utility scores in an inflammatory arthritis cohort. [References]. *Value in Health* 2011;921-7.

(49) Cauli A, Gladman DD, Mathieu A *et al.* Patient global assessment in psoriatic arthritis: a multicenter GRAPPA and OMERACT study. *J Rheumatol* 2011;**38**:898-903.

(50) Leung YY, Zhu TY, Tam LS *et al.* Minimal important difference and responsiveness to change of the SF-36 in patients with psoriatic arthritis receiving tumor necrosis factor-alpha blockers. *J Rheumatol* 2011;**38**:2077-9.

(51) Mease PJ, Woolley JM, Bitman B *et al.* Minimally important difference of Health Assessment Questionnaire in psoriatic arthritis: relating thresholds of improvement in functional ability to patient-rated importance and satisfaction. *J Rheumatol* 2011;**38**:2461-5.

(52) Davis AM, Palaganas MP, Badley EM *et al.* Measuring participation in people with spondyloarthritis using the social role participation questionnaire. *Ann Rheum Dis* 2011;**70**:1765-9.

(53) Leung Y-Y, Ho K-W, Zhu T-Y *et al.* Construct validity of the modified numeric rating scale of patient global assessment in psoriatic arthritis. *Journal of Rheumatology* 2012;**39**:844-8.

(54) Leung Y-Y, Png M-E, Wee H-L *et al.* Comparison of EuroQol-5D and short form-6D utility scores in multiethnic Asian patients with psoriatic arthritis: A cross-sectional study. *Journal of Rheumatology* 2013;**40**:859-65.

(55) Wink F, Arends S, McKenna SP *et al.* Validity and reliability of the Dutch adaptation of the Psoriatic Arthritis Quality of Life (PsAQoL) Questionnaire. *PLoS ONE* 2013;**8**:e55912.

(56) Coaccioli S, Bruno AA, Celi G *et al.* Validation of an original questionnaire for patients with psoriatic arthritis: the Psoriatic Arthritis Impact Profile (PAIP). *Clin Ter* 2014;**165**:e100-e108.

(57) Osterhaus JT, Purcaru O. Discriminant validity, responsiveness and reliability of the arthritis-specific Work Productivity Survey assessing workplace and household productivity in patients with psoriatic arthritis. *Arthritis Res Ther* 2014;**16**:R140.

(58) Gossec L, de WM, Kiltz U *et al.* A patient-derived and patient-reported outcome measure for assessing psoriatic arthritis: elaboration and preliminary validation of the Psoriatic Arthritis Impact of Disease (PsAID) questionnaire, a 13-country EULAR initiative. *Ann Rheum Dis* 2014;**73**:1012-9.

(59) Torre-Alonso JC, Gratacos J, Rey-Rey JS *et al.* Development and validation of a new instrument to measure health-related quality of life in patients with psoriatic arthritis: the VITACORA-19. *J Rheumatol* 2014;**41**:2008-17.

(60) Katchamart W, Benjamanukul S, Chiowchanwesawakit P. Validation of the Thai version of the Health Assessment Questionnaire for patients with psoriatic arthritis. *Int J Rheum Dis* 2014;**17**:181-5.

(61) Lebwohl M, Swensen AR, Nyirady J *et al.* The Psoriasis Symptom Diary: development and content validity of a novel patient-reported outcome instrument. *Int J Dermatol* 2014;**53**:714-22.

(62) Chiricozzi A, Bianchi L, Zangrilli A *et al.* Quality of life of psoriatic patients evaluated by a new psychometric assessment tool: PsoDisk. *European Journal of Dermatology* 2015;**25**:64-9.

(63) Lubrano E, Perrotta FM, Parsons WJ *et al.* Patient's global assessment as an outcome measure for psoriatic arthritis in clinical practice: A surrogate for measuring low disease activity? *Journal of Rheumatology* 2015;**42**:2332-8.

(64) Talli S, Etcheto A, Fautrel B *et al.* Patient global assessment in psoriatic arthritis - what does it mean? An analysis of 223 patients from the Psoriatic arthritis impact of disease (PsAID) study. *Joint Bone Spine* 2015.

(65) Leeb BF, Haindl PM, Brezinschek HP *et al.* Patient-centered psoriatic arthritis (PsA) activity assessment by Stockerau Activity Score for Psoriatic Arthritis (SASPA). *BMC Musculoskelet Disord* 2015;**16**:73.

(66) Naegeli AN, Flood E, Tucker J *et al.* The Worst Itch Numeric Rating Scale for patients with moderate to severe plaque psoriasis or psoriatic arthritis. *Int J Dermatol* 2015;**54**:715-22.

(67) Wilson HD, Mutebi A, Revicki DA *et al.* Reliability and validity of the psoriasis symptom inventory in patients with psoriatic arthritis. *Arthritis Care Res (Hoboken)* 2015.

(68) de Wit MP, Kvien TK, Gossec L. Patient participation as an integral part of patient-reported outcomes development ensures the representation of the patient voice: a case study from the field of rheumatology. *RMD Open* 2015;**1**:e000129.

(69) Tander B, Ulus Y, Terzi Y *et al.* Reliability and validity of the Turkish adaptation of VITACORA-19 in patients with psoriatic arthritis. *Archives of Rheumatology* 2016;**31**:2016.

(70) Piaserico S, Gisondi P, Amerio P *et al.* Validation and Field Performance of the Italian Version of the Psoriatic Arthritis Screening and Evaluation (PASE) Questionnaire. *Acta Derm Venereol* 2016;**96**:96-101.

(71) Leung YY, Thumboo J, Rouse M *et al.* Adaptation of Chinese and English versions of the Psoriatic Arthritis Quality of Life (PsAQoL) scale for use in Singapore. *BMC Musculoskelet Disord* 2016;**17**:432.

(72) Salaffi F, Di CM, Carotti M *et al.* The Psoriatic Arthritis Impact of Disease 12-item questionnaire: equivalence, reliability, validity, and feasibility of the touch-screen administration versus the paper-and-pencil version. *Ther Clin Risk Manag* 2016;**12**:631-42.

(73) Di CM, Becciolini A, Lato V *et al.* The 12-item Psoriatic Arthritis Impact of Disease Questionnaire: Construct Validity, Reliability, and Interpretability in a Clinical Setting. *J Rheumatol* 2016.

(74) Cohen JM, Halim K, Joyce CJ *et al.* Shedding Light on the "Hidden Psoriasis": A Pilot Study of the Inverse Psoriasis Burden of Disease (IPBOD) Questionnaire. *J Drugs Dermatol* 2016;**15**:1011-6.

(75) Cooper A, Wallman JK, Gulfe A. What PASSes for good? Experience-based Swedish and hypothetical British EuroQol 5-Dimensions preference sets yield markedly different point estimates and patient acceptable symptom state cut-off values in chronic arthritis patients on TNF blockade. *Scand J Rheumatol* 2016;**45**:470-3.

(76) Mease PJ. Measures of psoriatic arthritis: Tender and Swollen Joint Assessment, Psoriasis Area and Severity Index (PASI), Nail Psoriasis Severity Index (NAPSI), Modified Nail Psoriasis Severity Index (mNAPSI), Mander/Newcastle Enthesitis Index (MEI), Leeds Enthesitis Index (LEI), Spondyloarthritis Research Consortium of Canada (SPARCC), Maastricht Ankylosing Spondylitis Enthesis Score (MASES), Leeds Dactylitis Index (LDI), Patient Global for Psoriatic Arthritis, Dermatology Life Quality Index (DLQI), Psoriatic Arthritis Quality of Life (PsAQOL), Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F), Psoriatic Arthritis Response Criteria (PsARC), Psoriatic Arthritis Joint Activity Index (PsAJAI), Disease Activity in Psoriatic Arthritis (DAPSA), and Composite Psoriatic Disease Activity Index (CPDAI). *Arthritis Care Res (Hoboken)* 2011;**63 Suppl 11**:S64-S85.

(77) Orbai AM, Ogdie A. Patient-Reported Outcomes in Psoriatic Arthritis. *Rheum Dis Clin North Am* 2016;**42**:265-83.

(78) Tugwell P, Boers M, D'Agostino MA *et al.* Updating the OMERACT filter: implications of filter 2.0 to select outcome instruments through assessment of "truth": content, face, and construct validity. *J Rheumatol* 2014;**41**:1000-4.

(79) Garratt AM, Lochting I, Smedslund G *et al.* Measurement properties of instruments assessing self-efficacy in patients with rheumatic diseases. *Rheumatology (Oxford)* 2014;**53**:1161-71.

(80) Hendrikx J, de Jonge MJ, Fransen J *et al.* Systematic review of patient-reported outcome measures (PROMs) for assessing disease activity in rheumatoid arthritis. *RMD Open* 2016;**2**:e000202.

(81) Bartlett SJ, Orbai AM, Duncan T *et al.* Reliability and Validity of Selected PROMIS Measures in People with Rheumatoid Arthritis. *PLoS One* 2015;**10**:e0138543.

(82) Strand V, Crawford B, Singh J *et al.* Use of "spydergrams" to present and interpret SF-36 health-related quality of life data across rheumatic diseases. *Ann Rheum Dis* 2009;**68**:1800-4.

(83) Martin ML, McCarrier KP, Chiou CF *et al.* Early development and qualitative evidence of content validity for the Psoriasis Symptom Inventory (PSI), a patient-reported outcome measure of psoriasis symptom severity. *J Dermatolog Treat* 2013;**24**:255-60.

(84) Bushnell DM, Martin ML, McCarrier K *et al.* Validation of the Psoriasis Symptom Inventory (PSI), a patient-reported outcome measure to assess psoriasis symptom severity. *J Dermatolog Treat* 2013;**24**:356-60.

(85) Strober BE, Nyirady J, Mallya UG *et al.* Item-level psychometric properties for a new patient-reported psoriasis symptom diary. *Value Health* 2013;**16**:1014-22.

(86) Strober B, Zhao Y, Tran MH *et al.* Psychometric validation of the Psoriasis Symptom Diary using Phase III study data from patients with chronic plaque psoriasis. *Int J Dermatol* 2016;**55**:e147-e155.